



DELHI COLLEGE OF ENGINEERING

~~Q1~~ LIBRARY

CLASS NO.....310.510.....

BOOK NO.....KEN.....

ACCESSION NO.....41727.....

DATE DUE.

For each day's delay after the due date a fine of 3 nP. per Vol. shall be charged for the first week, and 25 nP. per Vol. per day for subsequent days.

[illegible]

MATHEMATICS OF STATISTICS

PART ONE

BY

J. F. KENNEY

University of Wisconsin

AND

E. S. KEEPING

University of Alberta

THIRD EDITION



(An East-West Edition)

D. VAN NOSTRAND COMPANY, INC.

TORONTO PRINCETON, NEW JERSEY
NEW YORK

LONDON

Affiliated East-West Press Private Ltd.

New Delhi

D. VAN NOSTRAND COMPANY, INC.
120 Alexander St., Princeton, New Jersey (*Principal office*)
24 West 40 Street, New York 18, New York

D. VAN NOSTRAND COMPANY, LTD.
358, Kensington High Street, London, W.14, England

D. VAN NOSTRAND COMPANY (Canada), LTD.
25 Hollinger Road, Toronto 16, Canada

Copyright © 1939, 1946, 1954, by

D. VAN NOSTRAND COMPANY, Inc.

Published simultaneously in Canada by
D Van Nostrand Company (Canada), Ltd.

No reproduction in any form of this book, in whole or in part (except for brief quotation in critical articles or reviews), may be made without written authorization from the publishers.

First Edition, 1939

Five Reprintings

Second Edition, 1946

Five Reprintings

Third Edition, April, 1954

Reprinted August 1956, July 1957,

November 1959, May 1961

Library of Congress Catalogue Card No. 53-13132

Affiliated East-West Press P Ltd.

East-West Student Edition — 1964

Reprinted in India with the special permission of the original publishers The D Van Nostrand Company Inc., Princeton, New Jersey, U.S.A. and the copyright holders

This book has been published with the assistance of the Joint Indian-American Standard Works Programme.

Published by W D Ten Broeck for AFFILIATED EAST-WEST PRESS PRIVATE LTD, C 57 Defense Colony, New Delhi 3 India, and printed by S M Balsaver at USHA PRINTERS, National House, Tulloch Road, Bombay 1, India

T.
Silver

PREFACE

The first edition of this book was issued in 1939, and a second edition appeared in 1946. Although some minor alterations were made at that time, and the text was clarified in places, the scope and character of the book remained essentially unchanged. It dealt with descriptive statistics and relegated almost all considerations of sampling and statistical inference to Part Two, where a mathematically more adequate treatment was possible.

Within comparatively recent years a great change has come over the teaching of elementary statistics. Whereas formerly the emphasis was on the technique of organizing and classifying the data supplied by observation or experiment so as to expose their essential characteristics, the tendency now is to stress the limitations of statistical inference and the uncertain nature of conclusions from observational data. Ever since W. S. Gosset, in 1908, gave a method of estimating the error in the mean of a small sample, it has indeed been known that the traditional statistical methods were inadequate, but the mathematical difficulties of any rigorous treatment of statistical inference kept this topic out of the elementary classroom. For some time it was thought that a sound treatment could not be given without using advanced mathematics, and it is still not easy to do so, but latterly it has been found possible to develop a satisfactory exposition of the principal techniques of inference, suitable to the student with a very limited mathematical background. The important thing is to know what assumptions are made in the mathematical treatment and to be able to judge whether these assumptions can reasonably be regarded as approximately fulfilled in the data under test.

The present edition (the third) represents a radical revision and extension of the text, in line with the ideas stated in the previous paragraph. There is still a need for descriptive statistics, and the first eight chapters deal with topics which may be included under this general heading. These chapters require little mathematics beyond elementary algebra. The notation of the definite integral has been used for convenience in Chapter VIII (on the Normal Law), but the simple interpretation as an area is all that is necessary.

Some notions of probability are essential for the understanding of statistical inference, and Chapter IX gives a treatment based on the rules of combination for relative frequencies. Prior knowledge of the elementary algebra of permutations and combinations is not assumed. The sections on continuous probability distributions use very simple notions of calculus (the integration of x^n for positive integral values of n), but these sections are not essential for later topics in this book.

Chapters X to XIII introduce the ideas of significance, of confidence levels, and of the testing of hypotheses, as applied to several important types of statistic such as the proportion of individuals in a sample having a certain characteristic, the sample mean for a measured variable, the difference of two sample means, the sample variance, and the ratio of two variances. Non-parametric tests (of goodness of fit and randomness) are discussed in Chapter XIII, as well as some small-sample tests for order statistics.

Chapters XIV to XVI, on time series, regression, and correlation, have been considerably expanded from the treatment in the earlier editions, to include tests of significance for regression and correlation coefficients, the use of chi-square for contingency tables, and Fisher's method for 2×2 tables.

With the idea of making the book more suitable than before for classes in business statistics, a chapter on Index Numbers has been added as an illustration of weighted averages, and the usual methods of analysis of time series have been described in Chapter XIV. However, topics which are clearly non-mathematical, such as the preparation of diagrams and statistical tables, or the precise wording of questionnaires, or the techniques of sampling in an actual survey, have been passed over very lightly indeed. It is extremely doubtful whether a first course in statistics is the place for an extended treatment of these topics.

Worked examples have been freely used throughout the text. Some of these have been chosen from Canadian statistical sources (one of the authors being a Canadian), but the great majority are American. Some of the sets of data used are artificial, but they illustrate the point at issue. An instructor who feels that it is important for his students to work on up-to-date material can probably find what he wants somewhere in the flood of recent official Government or United Nations statistical publications, or in the records of experimental work carried out by himself or his colleagues. We feel that an understanding of the method he is using is more important for the student of statistics than the particular data on which he practices.

The tables in the Appendix, and interspersed in the text, have been considerably increased in number, so that the student now has at hand all the tables necessary for the application of the common statistical tests. (The table of logarithms that was included in the earlier editions has been omitted, as such tables are readily available.) A copy of Barlow's *Tables of Squares, Cubes, Square Roots, Cube Roots and Reciprocals* (New York, Tudor Publishing Co., Inc.) is very convenient to have at hand when engaged in computing, and Fisher and Yates' *Statistical Tables for Biological, Agricultural and Medical Research* (Oliver and Boyd) is an invaluable reference. For permission to reprint tables or parts of tables, the authors are grateful to Sir Ronald Fisher (Tables II and III), Prof. G. W. Snedecor (Table IV), Dr. C. Eisenhart (Table 47), Dr. W. J. Dixon (Table 49), Dr. E. S. Pearson (Table V and Charts

I and II), Dr. P. C. Mahalanobis (Table 51), Dr. P. R. Rider (Table 53), and Dr. J. W. Campbell (Table VI).

The references at the ends of many of the chapters are intended to direct the attention of the elementary student to a few selected books or journal articles of not too technical a nature, from the reading of which he might be expected to profit. References are also given to a few articles from which tables or parts of tables have been quoted, so that the originals may be consulted in a good library; but no attempt has been made to give chapter and verse for all methods and formulas which are mentioned in the text. It would be impossible, even if it were desirable, to mention all those books, papers, and lectures from which the authors have derived inspiration and upon which they have perhaps leaned.

The authors are indebted to Professor A. T. Craig, State University of Iowa, and to Mr. Albert Shaw, formerly lecturer at the University of Alberta, and now Assistant Provincial Statistician, for reading the manuscript and for correcting several errors and dubious statements. For what errors and obscurities remain (and it is surely too much to hope that none remain!) the authors must accept the blame.

J. F. K.

E. S. K.

CONTENTS

CHAPTER	PAGE
INTRODUCTION	1
0.1 The Scope of Statistics. — 0.2 Mathematics and Statistics. — 0.3 Calculating Machines. — 0.4 Collateral Reading.	
I FREQUENCY DISTRIBUTIONS	5
1.1 Variables and Constants. — 1.2 Variates. — 1.3 Errors of Measurement. — 1.4 Accuracy. — 1.5 Significant Figures. — 1.6 Sources of Data. — 1.7 Classification and Tabulation. — 1.8 Frequency Distributions. — 1.9 Class Intervals. — 1.10 Class Limits and Class Boundaries. — 1.11 Cumulative Frequencies. — Exercises.	
II GRAPHICAL REPRESENTATION	21
2.1 The Function Concept. — 2.2 Charts. — 2.3 Frequency Polygons. — 2.4 Histograms. — 2.5 Frequency Curves. — 2.6 Cumulative Frequency Polygons. — 2.7 Ogive Curves. — Exercises.	
III THE MEDIAN AND OTHER QUANTILES	32
3.1 Averages. — 3.2 The Median. — 3.3 Quartiles. — 3.4 The Quartile Deviation. — 3.5 Quantiles. — 3.6 Percentile Ranks. — 3.7 Approximate Characterization of a Distribution by Quan- tiles. — Exercises.	
IV THE ARITHMETIC MEAN AND OTHER AVERAGES .	42
4.1 Various Averages. — 4.2 Notation for Sums and Products. — 4.3 Arithmetic Mean. — 4.4 Weighted Arithmetic Mean. — 4.5 Arithmetic Mean of a Grouped Variate. — 4.6 Mean of Means. — 4.7 The Mode. — 4.8 Relation between Mean, Med- ian, and Mode. — 4.9 Relative Merits of Mean, Median, and Mode. — 4.10 Geometric Mean. — 4.11 Weighted Geometric Mean. — 4.12 The Law of Growth. — 4.13 Harmonic Mean. — 4.14 Relation of Arithmetic, Geometric, and Harmonic Means. — 4.15 Root Mean Square. — Exercises.	
V INDEX NUMBERS	64
5.1 Index Numbers as Weighted Averages. — 5.2 Price Index Numbers. — 5.3 An Example. — 5.4 Quantity Index Num- bers. — 5.5 Fisher's Tests for Index Numbers. — 5.6 Geometric and Harmonic Means of Relatives. — 5.7 Series of Index Num- bers. — 5.8 Adjusted Death Rates. — Exercises.	

CHAPTER	PAGE
VI STANDARD DEVIATION AND OTHER MEASURES OF DISPERSION	75
6.1 Various Measures of Dispersion. — 6.2 The Range. — 6.3 Deviations. — 6.4 Mean Absolute Deviation. — 6.5 The Standard Deviation. — 6.6 Calculation of the Standard Deviation. — 6.7 Standard Deviation of a Grouped Continuous Variate. — 6.8 Charlier Check. — 6.9 Grouping Error of the Standard Deviation. — 6.10 Meaning of the Standard Deviation. — 6.11 Relative Dispersion. — 6.12 Some Theorems on Variance. — Exercises.	
VII MOMENTS. SKEWNESS AND KURTOSIS	90
7.1 Populations and Samples. — 7.2 Moments about the Origin. — 7.3 Moments about the Mean. — 7.4 Relations between the m_r' and the m_r . — 7.5 Calculation of Third and Fourth Moments. — 7.6 Sheppard's Corrections for Grouping Errors. — 7.7 Standard Units. — 7.8 Moments in Standard Units. — 7.9 The k -statistics. — 7.10 Skewness. — 7.11 Other Measures of Skewness. — 7.12 Kurtosis. — 7.13 Specimen Computation of Moments. — Exercises.	
VIII THE NORMAL CURVE	107
8.1 Frequency Curves. — 8.2 The Normal Curve. — 8.3 Standard Form. — 8.4 Tables of Ordinates and Areas. — 8.5 Properties of the Normal Curve. — 8.6 Fitting a Normal Curve to a Distribution. — 8.7 Graduation. — 8.8 Justification for Using the Normal Curve. — 8.9 Purpose of Graduating a Curve. — 8.10 Normalizing an Ordered Series. — Exercises.	
IX PROBABILITY	124
9.1 Meaning of Probability. — 9.2 Combination of Relative Frequencies. — 9.3 Rules for Combining Probabilities. — 9.4 Permutations. — 9.5 Combinations. — 9.6 Some Problems in Probability. — 9.7 Simple and Compound Events. — 9.8 Continuous Probability. — 9.9 Moments of a Probability Distribution. — 9.10 Mathematical Expectation. — 9.11 Statistics and Probability. — 9.12 Mathematical Models. — Exercises.	
X THE BINOMIAL AND POISSON DISTRIBUTIONS . . .	144
10.1 A Coin-tossing Problem. — 10.2 Binomial Coefficients. — 10.3 The Binomial Distribution. — 10.4 Moments of the Binomial Distribution. — 10.5 Fitting a Binomial to a Given Distribution. — 10.6 The Poisson Distribution. — 10.7 Moments of the Poisson Distribution. — 10.8 Fitting a Poisson Distribution. — 10.9 Poisson Distribution for Random Events. — Exercises.	

CHAPTER	PAGE
XI SIGNIFICANCE TESTS FOR BINOMIAL POPULATIONS	161
11.1 Approximation of the Binomial by a Normal Distribution. — 11.2 Significance of an Observed Proportion — 11.3 Tests of Hypotheses. — 11.4 Confidence Limits for the Binomial Parameter. — 11.5 Confidence Interval Charts. — 11.6 Mean and Variance of a Linear Combination of Independent Variates. — 11.7 Significance of a Difference Between two Sample Proportions. — 11.8 Confidence Limits for the Difference in Proportions. — 11.9 Binomial Probability Paper. — 11.10 Sampling from a Finite Population. — Exercises.	
XII SIGNIFICANCE OF MEANS AND VARIANCES	175
12.1 Distribution of the Sample Mean. — 12.2 An Illustration of the Distribution of Means. — 12.3 Significance of Means. — 12.4 Confidence Interval for Means. — 12.5 Distribution of the Sample Sum. — 12.6 Correction for Finite Population. — 12.7 Significance of Difference of Means in Large Samples. — 12.8 Student's <i>t</i> -Distribution. — 12.9 Degrees of Freedom. — 12.10 Confidence Limits for the Mean, for Small Samples. — 12.11 Confidence Limits for the Difference of Means, for Small Samples. — 12.12 Significance of Differences in Paired Samples. — 12.13 Distribution of the Sample Variance. — 12.14 Significance of a Ratio of Two Variances — The <i>F</i> -Distribution. — 12.15 Test for Homogeneity of Variance. — 12.16 Analysis of Variance. — 12.17 Control Charts. — Exercises	
XIII NON-PARAMETRIC AND ORDER STATISTICS	197
13.1 Non-parametric Statistics. — 13.2 Goodness of Fit. — 13.3 Pooling of Class Frequencies. — 13.4 The Chi-Square Test of Hypotheses. — 13.5 The Chi-Square Test of Goodness of Fit for Graduated Distributions. — 13.6 Tests of Randomness. — 13.7 Distribution of Number of Runs — 13.8 Run Test of Difference Between Two Samples — 13.9 Random Numbers. — 13.10 The Sign Test for Differences in Paired Samples. — 13.11 Inequalities of the Tchebycheff Type. — 13.12 Order Statistics. — 13.13 The Median. — 13.14 Estimation by Percentiles. — 13.15 The Range. — 13.16 Quotient of Ranges in Samples from a Rectangular Population. — Exercises.	
XIV TIME SERIES	219
14.1 Time as a Variable. — 14.2 Moving Averages. — 14.3 The Slutsky-Yule Effect. — 14.4 Mathematical Trend Lines. — 14.5 Linear Functions. — 14.6 Fitting a Straight Line. — 14.7 Graphical Method. — 14.8 Method of Moments. — 14.9 The Method of Least Squares. — 14.10 Fitting a Straight Line Through the Origin. — 14.11 Simplification of Calculations for Equispaced Data. — 14.12 Exponential Trends. — 14.13 The Compound Interest Law. — 14.14 Semi-logarithmic Graph Paper. — 14.15 Ratio Charts. — 14.16 Logarithmic Graph	

CHAPTER	PAGE
Paper. — 14.17 Other Types of Trend. — 14.18 The Analysis of Business Time Series — 14.19 Deflating and Deseasonalizing Data. — 14.20 Elimination of Trend and Irregularities. — Exercises.	
XV LINEAR REGRESSION AND CORRELATION	252
15.1 Bivariate Data. — 15.2 Regression. -- 15.3 Coefficient of Correlation. — 15.4 Relation between Coefficients of Regression and Correlation -- 15.5 Interpretation of the Coefficient of Correlation -- 15.6 Variation Around the Regression Line. — 15.7 Significance of the Regression Coefficients -- 15.8 Significance of the Correlation Coefficient -- 15.9 Accuracy of Estimate from Regression. -- 15.10 Calculation of r for Grouped Variates — 15.11 Regression Lines for a Correlation Table. — 15.12 Variance of Estimate for a Correlation Table. — 15.13 Normal Correlation Surface -- 15.14 Best-fitting Straight Line When Both Variates Are Subject to Error. — 15.15 Which Regression Should Be Used for Prediction? -- Exercises.	
XVI FURTHER TOPICS IN CORRELATION	286
16.1 Reliability and Validity of Tests — 16.2 Analysis of Variance of Test Scores -- 16.3 Rank Correlation -- 16.4 Parabolic Regression — 16.5 Correlation Index for Non-linear Regression. — 16.6 Curves of Column and Row Means -- 16.7 Calculation of Correlation Ratios for Grouped Variates. — 16.8 Test for Linearity of Regression with Grouped Variates. — 16.9 Some General Remarks on Correlation. -- 16.10 Contingency. — 16.11 Chi-square Test for Association. -- 16.12 Coefficient of Contingency. — 16.13 The 2×2 Table. -- 16.14 Yates' Correction — 16.15 Fisher's Exact Method for 2×2 Tables. — 16.16 Problems Involving Three Variates. — Exercises.	
REVIEW QUESTIONS AND PROBLEMS	313
APPENDIX:	
Table I Ordinates and Areas of the Normal Curve . . .	319
II Values of t Corresponding to Given Probabilities .	322
III Values of χ^2 Corresponding to Given Probabilities	324
IV 5% and 1% points for the Distribution of F . .	326
V Random Sampling Numbers	330
VI Values of $\tanh z'$	334
Chart I Confidence Limits (95%) for the Binomial Distribution	338
Chart II Confidence Limits (95%) for the Correlation Coefficient	339
ADDITIONAL ANSWERS	340
INDEX	343

FOREWORD

This book, in its present form, has been used as the basis for a 3-hour course (running for about 26 weeks) at the University of Alberta. Students come from the Faculties of Arts and Science and Education and have different mathematical backgrounds, but many of them have only high school algebra. It has been found possible to cover most of the material, with the exception of §§12.15, 12.16, 13.10–13.16, 14.17, 15.12–15.14, 16.1, 16.2, 16.8, and 16.16, which are either passed over lightly or omitted altogether. In the usual two-semester course at an American university, which is rather longer than the Alberta year, some of this omitted material might be included.

It is the practice at Alberta to combine a 3-hour-a-week laboratory period with the lecture course. During these periods actual computations are carried out, with the help of hand-operated calculating machines (the Monroe Educator model) and reports are subsequently written up. If a laboratory is not feasible, the students should at least work a considerable number of the numerical exercises at the ends of the chapters in order to gain facility in the handling of data.

A fuller treatment of some important topics, such as the Analysis of Variance, which are touched on only lightly in this book, may be found in Part Two of our *Mathematics of Statistics*, Van Nostrand, 2nd edition, 1951.

MATHEMATICS OF STATISTICS

INTRODUCTION

0.1 The Scope of Statistics. As the name implies, statistics originally meant information useful to the *state*, for such purposes as taxation or the raising of an army. Later, it came to mean quantitative data which tend to fluctuate in a more or less unpredictable way, and it is still used popularly in this sense, as when the newspapers talk about the statistics of highway accidents or of births, marriages, and deaths.

More recently, statistics has usually meant the science (and art) concerned with the collection, presentation, and analysis of quantitative data so that intelligent judgments may be formed upon them, as exemplified, for instance, in the statistical reports presented by many large companies to their shareholders. For a great many practicing statisticians this is, in fact, the most important part of their work, the results of which are embodied in neat tables and diagrams. But although many difficult problems arise in the collection and processing of the raw data, these problems are not usually of a mathematical nature, and a good deal of the work is of a routine character:

Of late years statistics has more and more come to the help of the other sciences, particularly the biological and social sciences, as an aid to the intelligent planning of experiments (so as to secure the maximum of information for a given expenditure of time and money) and as a means of assessing the significance of the results obtained by experiment. In the "exact" sciences, such as physics and astronomy, with their relatively high precision of measurements, there did not appear to be the same need for statistical methods as in agriculture, medicine, economics, and many other fields, where the results of experiments are complicated by a multiplicity of factors beyond the control of the observer. For a long time, in fact, statistics in the physical sciences hardly progressed beyond the calculation of a standard error (or, more likely, a "probable error") and the occasional fitting of a curve by least squares. There is now, however, a keen appreciation among physicists of the fundamental role of statistics in the treatment of the complex problems of molecular, atomic, and nuclear structure.

Moreover, in a very different field, statistics has invaded industry, and statistical control is now applied by many large manufacturing concerns to ensure that the quality of their product remains reasonably constant. It is broadly true, however, that the bulk of modern applications of statistical methods is in the biological, psychological, and sociological sciences.

From the point of view of these and other sciences, statistics may be

regarded as the technique of drawing valid conclusions from a limited body of experimental or observational data. Occasionally, as in a decennial census of the whole population of a country, the results of observation may be regarded for practical purposes as exact, but such a census is extremely costly and time-consuming; as a rule, inferences must be made about a population on the basis of observations made on a few relatively small samples. Such inferences are not certain; they are merely more or less probable, and the methods of modern statistics enable us to estimate the probability of any conclusions which may be drawn.

0.2 Mathematics and Statistics. The proper treatment of observational data is primarily the concern of the observer who collects them. A large literature has grown up describing appropriate procedures for estimating the magnitude of observed effects and testing a variety of hypotheses regarding them. This body of material comprises what is known as Experimental Statistics. Usually, however, the procedures and tests of experimental statistics are not really valid unless the observations satisfy some rather stringent conditions, which they may easily fail to do. The discussion of the exact conditions under which these procedures are valid, and the development of new kinds of tests and new designs of experiments, are the concern of Mathematical Statistics. This subject is now a discipline of its own, recognized as a distinct department in several universities. It is a highly specialized subject, which utilizes advanced and rather abstract mathematical ideas and is therefore well beyond the level of this book. Accordingly, no attempt will be made here to justify the use of the various tests described in the chapters dealing with statistical inference. Proofs of some of these will be found in Part Two*, which is more advanced mathematically.

0.3 Calculating Machines.† A full description of the parts of a calculating machine and their operation may be obtained from an *Instruction Book* furnished by the manufacturer, so only a brief description will be given here.

A calculating machine is constructed to add and subtract. By means of continued addition or subtraction, operations involving multiplication, division, and square root can also be performed with great speed.

In addition to a keyboard on which numbers can be punched, most machines have a sliding carriage, carrying two dials one above the other. These dials are called *revolution register* (upper dial) and *product register* (lower dial). In finding a product nx , one of the factors n is punched on the keyboard and as the motive crank at the side is turned, the other factor x appears on the upper dial. The product nx is then read from the lower dial.

An important property of the modern calculating machine is its adapt-

* Kenney and Keeping, *Mathematics of Statistics*, Part Two, 2nd Ed., D. Van Nostrand Co., Inc., New York, 1951.

† The early history of modern computing machines is outlined in the *American Mathematical Monthly*, **31** (1924), pp. 422-429.

ability to short cuts and combinations of operations. For example, one may multiply two numbers nx and add the result to a third number k without tabulating the intermediate steps. This is accomplished by punching the number k on the keyboard, transferring it to the lower dial (product register), and then proceeding as in finding the product nx . The result $nx + k$ is then read from the lower dial. An extension of this procedure is especially useful in a series of computations where k and n are constant and various values are assigned to x . To describe the procedure, suppose it is required to calculate the successive values of $12 + 6x$ for $x = 5, 7, 15, 12$, etc. The number $k = 12$ is first registered on the lower dial, then the factor $n = 6$ is placed on the keyboard, and by turning the crank forward five times to make the first value of $x = 5$ appear on the upper dial, the result $12 + 6 \times 5$ appears on the lower dial. Instead of clearing the dial, the crank is now turned forward twice more to rebuild the value $x = 5$ into $x = 7$, and the result $12 + 6 \times 7$ can be read from the lower dial. In rebuilding $x = 15$ into $x = 12$ the crank is turned backwards. This procedure can be repeated until all the required values of $12 + 6x$ have been calculated. A process of this sort is called the *continuous method* of calculating.

In most of the exercises in this course, the computations are not particularly laborious and calculating machines are not really necessary. However, if machines are available they will be found a great help. Elaborate, fully automatic, electrically operated machines are not required; the small hand-operated models, such as the Monroe Educator, are quite good enough for practice. The trained statistician, of course, will profit by all the devices with which the modern high-speed machine is equipped.

0.4 Collateral Reading. Perhaps no single textbook can meet all the needs of all students of statistics. There are several good books on elementary statistics which, although not fundamentally different, present different points of view on certain topics and treat them with varying degrees of emphasis depending upon the field of major interest. At least some of the books listed below should be readily available on the reserve shelf of the library. The list should be useful to those who wish to study more fully certain details in which they may be interested.

1. F. E. Croxton, and D. J. Cowden, *Practical Business Statistics* (Prentice-Hall, Inc., 1948). Mainly nonmathematical, with emphasis on the procedures customary in business. Contains some useful tables.

2. F. E. Croxton, and D. J. Cowden, *Applied General Statistics* (Prentice-Hall, Inc., 1939). An encyclopedic collection of statistical methods and elementary theory. Very useful to the practicing statistician.

3. W. E. Deming, *Statistical Adjustment of Data* (John Wiley & Sons, Inc., 1943). Concerned primarily with curve fitting by the method of least squares.

4. W. E. Deming, *Some Theory of Sampling* (John Wiley & Sons, Inc., 1950). A discussion of the techniques and errors of sampling, with some general mathematical theory of the design and analysis of sampling procedures.

5. W. J. Dixon, and F. J. Massey, *An Introduction to Statistical Analysis* (McGraw-Hill Book Co., Inc., 1951). Modern procedures, illustrated by examples from many fields, but without mathematics beyond very elementary algebra. Contains a good selection of tables.

6. C. H. Gouldeen, *Methods of Statistical Analysis* (John Wiley & Sons, Inc., 1951). Practical methods of value to the agronomist, plant breeder etc. Intended for those who already know some elementary statistics.

7. E. L. Grant, *Statistical Quality Control* (McGraw-Hill Book Co., Inc., 1952). Explains the methods and objects of the statistical control of fluctuation in the quality of output of factory products. Nonmathematical.

8. P. J. Hoel, *Introduction to Mathematical Statistics* (John Wiley & Sons, Inc., 1947). Covers much ground briefly, for students with at least a year of calculus.

9. P. O. Johnson, *Statistical Methods in Research* (Prentice-Hall, Inc., 1949). Applies modern statistical methods to the interpretation of observations, particularly in the field of education. Some mathematical proofs are given but mainly as appendices.

10. F. C. Mills, *Statistical Methods* (Henry Holt & Co., 1938). A well-known book on the applications to business and economics. Traditional material.

11. M. J. Moroney, *Facts From Figures* (Penguin Books, 1951). A popular presentation, written in a lively style, of practically the whole field of elementary statistics.

12. J. Neyman, *First Course in Probability and Statistics* (Henry Holt & Co., 1950). Intended for students with some mathematical maturity. The treatment is based on probability and illustrated by problems in genetics.

13. R. Pearl, *Introduction to Medical Biometry and Statistics* (W. B. Saunders Co., 1930). Deals with the collection and presentation of data, and simple curve fitting, in the realm of biometry. Now rather out-of-date, but a classic.

14. G. W. Snedecor, *Statistical Methods* (Iowa State College Press, 1946). An authoritative account of methods used in agriculture, with many examples worked in detail. Uses only elementary algebra.

15. L. H. C. Tippett, *Methods of Statistics* (Williams and Norgate, 1941). A good practical exposition of R. A. Fisher's techniques, useful particularly to the research worker in biology.

16. H. M. Walker, *Studies in the History of Statistical Method* (The Williams and Wilkins Co., 1929). For the historically-minded student, gives a fascinating account of the origin of several important statistical concepts. Contains a bibliography of the earlier literature.

17. H. M. Walker, *Elementary Statistical Methods* (Henry Holt & Co., 1943). A clear treatment of elementary statistics, with an introduction to sampling, requiring only high-school mathematics, but emphasizing symbolism.

18. H. M. Walker, and J. Lev., *Statistical Inference* (Henry Holt & Co., 1953). A simple and relatively non-mathematical introduction to the modern concepts of statistical inference and the testing of hypotheses.

19. C. E. Weatherburn, *A First Course in Mathematical Statistics* (Cambridge University Press, 1946). A clear and compact treatment of the mathematics underlying common statistical procedures. Written at about the level of Part Two of this work.

20. S. S. Wilks, *Elementary Statistical Analysis* (Princeton University Press, 1948). A freshman course, presupposing a semester of elementary mathematical analysis, including a little calculus. Strongly emphasizes probability, statistical significance, and confidence limits.

21. H. H. Wolfenden, *The Fundamental Principles of Mathematical Statistics* (Actuarial Society of America, 1942). Presents those parts of the theory needed by actuaries and workers in the field of vital statistics, including the graduation of data by theoretical curves.

22. G. U. Yule, and M. G. Kendall, *An Introduction to the Theory of Statistics*, Fourteenth Edition (C. Griffin, 1950). A standard work, which has been revised several times. An excellent account of some of the more theoretical aspects of statistics.

CHAPTER I.

FREQUENCY DISTRIBUTIONS

1.1 Variables and Constants. A *variable* is a quantity which may take on any value from a given set of values, called its *domain*. Thus, the sex of an animal is a variable of which the domain contains the two values, male and female. The daily rainfall at a certain place is a variable of which the domain includes all rainfalls from zero up to some indefinite upper limit which is the largest daily rainfall conceivable.

A variable whose domain contains only one value (in a particular situation or discussion) is called a *constant*. A constant which changes its value from one situation to another is often called a *parameter*. In the equation $y = 5x + c$, representing a family of parallel straight lines, the c is a parameter. Its value changes from one line to another, but is constant (say 3) for all points (x, y) on one given line.

Statistics is concerned with variables that fluctuate in a more or less unpredictable way, such as the monthly total of highway accidents in the state of New York or the daily yield of milk by a certain cow. These are said to be *random variables*. The particular day of the week corresponding to a date in the future, such as June 17, 2049, is not a random variable because a rule can be given for calculating it exactly, assuming that the present Gregorian calendar persists. There may be assignable causes with predictable effects on the total of highway accidents in a certain month, but no one would expect to be able to predict this total exactly. The essence of a random variable is that some part of it is unpredictable.

1.2 Variates. The raw material of statistics consists of numbers usually obtained by some process of counting or measurement. These are referred to collectively as the *data*. It is convenient to replace the values of a random variable by numbers. These numbers may be assigned in a rather arbitrary way, as when we denote a female by 0 and a male by 1, or they may be the numerical value, in suitable units, of a measurement, as when we represent the height of an individual by a certain number of inches or centimeters. In both instances we replace our variable by a *variate*, the domain of which is always a set of real numbers, and which can be denoted by a letter such as x or y . Thus, if the variable is the face that comes to rest uppermost when an ordinary die is rolled, the corresponding variate is naturally the number of spots on this face, with a domain consisting of the real numbers 1, 2, 3, 4, 5, 6.

It is evident from the foregoing illustrations that variates are of two kinds: *discrete* and *continuous*. Discrete variates have domains restricted to isolated

real numbers, most frequently positive integers or zero. Examples are the number of children in a family and the number of heads in ten tosses of a coin, where the values are obtained by *counting*. Continuous variates correspond to variables which are *measured*, theoretically to any degree of fineness. Such variables, for instance, are height, weight, and temperature. The domain of a continuous variate is an interval, or set of intervals, on the real number axis.

In some investigations the variable cannot be accurately measured, but individuals can be *ranked* more or less accurately. Thus a foreman may rank a number of his workmen in order of competence or a judge in a taste trial may rank several varieties of ice cream in order of preference. The rather vaguely defined variable is thus replaced for statistical purposes by a discrete variate with domain consisting of the numbers 1, 2, 3, . . .

The concept of randomness, as applied to a variable, means that to every value of the corresponding variate within its domain we can assign a *probability*, namely, the probability that (in a given set of circumstances) that particular value will actually be realized if a measurement or count is made.* The term "probability" will be discussed in more detail later on, but for the present we assume that its meaning is at least vaguely understood and that it is a number between 0 and 1 inclusive, an impossible value having the probability 0 and one that is absolutely certain the probability 1. If we toss an ordinary coin repeatedly we know from experience that about half the time it will come down heads and about half the time tails. We may think of a variate x with the values 0 (for heads) and 1 (for tails). It seems reasonable to allot to each of these a probability approximately $\frac{1}{2}$, since we know that both are about equally likely and one or the other of them must occur in any toss.† In the same way, the probabilities for a well-made die corresponding to each of the six possible values of x must be nearly $\frac{1}{6}$, since we have good reason to believe that each of the six faces is as likely to show up as any other and it is certain that one of them will.

If the variable is height, say of an adult male American, and the corresponding variate is the number of inches in the height of an individual, its domain is a continuous interval extending from, say, 50 to 85. The probabilities are, however, relatively extremely small near either end of the domain, since dwarfs and giants are rare, and relatively large between 65 and 75, since most adult males have heights somewhere in this range.

1.3 Errors of Measurement. If fifty boys each try to measure the length of a classroom, along the floor between the two end walls, to the nearest sixteenth of an inch, using the same ordinary 12-inch ruler, they will certainly come up with a variety of answers. These answers will probably cluster

* The terms "random variable," "chance variable," and "stochastic variable" are often used in the literature as synonyms of "variate" in the sense given above.

† Cases in which the coin stands on end are ignored!

around an average value, with a few values scattering rather widely. If we assume for the present that the average value is a good approximation to the true value, the differences of the various values from the average represent errors of measurement. These errors constitute the random element in the measurement of the length. They are due to many causes, partly psychological, but have two important characteristics: positive and negative errors tend to occur about equally often and small errors are much more probable than large ones. Of course, gross errors, such as might be produced by reading a 3 on a scale as an 8, are excluded from consideration.

Another kind of error may be due to imperfection in the measuring instrument itself, and this error is called *systematic* or *biased*, as opposed to random. Thus it may well happen that the 12-inch ruler used to measure the classroom is itself one-sixteenth of an inch short, and then all the measurements made with it will be too large. The errors are all in the same direction and do not tend to cancel when an average is taken. It has been found in accurate astronomical work, when observers have to time the exact instant of passage of a star-image over the crosshairs of a fixed telescope, that many people have systematic personal errors, tending always to be a little fast or a little slow in their reactions. Once this has been determined it can be allowed for, and, in this particular type of observation, personal error can be eliminated altogether by using suitable photoelectric devices instead of the human eye. It is the aim of good experimental work to eliminate systematic errors as far as possible, but there practically always remains a residuum of unavoidable error which is assumed to be random.

If the measurement consists solely of counting a few well-defined objects, there is, of course, no error. If a man gives the number of his children to the census-taker as 6, that figure may be assumed exact. Yet if the total population of the United States is given as 150,697,361 (the census figure for April 1, 1950) this number, even though obtained by counting, is almost certainly not exact. There are too many possible sources of error (people not counted or counted twice, or born just after April 1 and counted, etc.) for us to believe that the figure was correct even on the day of the census.

1.4 Accuracy. We have seen that most statistical data relate to variates which are subject to random error and therefore are only approximately correct as measured. With continuous variates, the observed values as recorded can never be absolutely established by measurement. Thus, the height or weight of an object can be measured only approximately, the error depending upon the precision of the instrument and the care and accuracy of the observer. However, it is not always necessary that measurements be recorded as accurately as it is possible to make them. Similarly, with discrete variates the standard of accuracy used may be less than it is possible to obtain. In population statistics, for example, it may be sufficient to record the numbers to

the nearest thousand, with three zeros at the end to fill out to the decimal point. Thus,

<i>City</i>	<i>Population</i>
A	326,000
B	729,000

On the other hand, the exact number of students in a university might be required. The degree of accuracy needed is determined by the purpose of the investigation and it is limited by the closeness with which the variables can be measured.

It follows, therefore, that the degree of accuracy in the final result of a problem involving computations is limited by that of the original data. Students sometimes carry results of problems to five or more decimal places when the original data do not justify more than two or three decimal places. A table of measurements which constitutes the raw data for a statistical investigation should always specify the degree of accuracy in the readings. Thus, if monthly rainfall is being measured to the nearest hundredth of an inch, and one measurement seems to be exactly 5 in., it should be recorded as 5.00 in., with two zeros. A measurement that is merely recorded as 5 means it is correct to the nearest integer and its true value lies between 4.5 and 5.5, whereas 5.00 means the true value is known to lie between 4.995 and 5.005. The three digits in 5.00 are said to be significant.

1.5 Significant Figures. A clear understanding of the meaning of significant figures is important in numerical work. In a number recorded as the result of a measurement, all digits except zero are always significant. Zeros which commence a number are nonsignificant; they are merely position-fillers necessary to indicate the meaning of the first significant digit. Thus in 0.00327 there are only three significant figures, and the 3 means 3 thousandths. Zeros which conclude a number are significant if they follow the decimal point, as in 5.200, but if they lie to the left of the decimal point they may or may not be significant. It is impossible to tell merely by looking at the number 186,000 whether there are three, four, five, or six significant figures in it. If the number is written in scientific notation as 1.86×10^5 , we know, however, that only three figures are significant. To indicate that six figures are significant we should write it as 1.86000×10^5 , and this would mean that the number lies between 185,999.5 and 186,000.5. Zeros which occur elsewhere than at the beginning or the end are always significant, as in 1.002.

In multiplying or dividing numbers time is saved, and a delusive appearance of accuracy is avoided, if the numbers with most significant figures are first rounded off. If no number contains fewer than s significant figures, the others should be rounded off, if necessary, to $s + 1$ figures. After the calculation has been performed, the result is rounded off to s figures. Thus, if we wish to evaluate $v = (4/3)\pi r^2 h$, where $r = 22.264$ and $h = 7.2$, these being measured

values, we round off r to 3 figures as 22.3. The number $4/3$ is exact and therefore not subject to error. The number π is also exact but can be approximately represented by a decimal with as many significant figures as required. Here it may be taken to 3 figures as 3.14. The calculation gives $v = 14,990$, which may be rounded off as 1.50×10^4 . According to the foregoing rule, only two significant figures should be kept, since h has only two figures, and this would mean writing $v = 1.5 \times 10^4$, but it is better to interpret the rule somewhat liberally. Remember that a 3-digit number beginning with 1 is not very different in magnitude from a 2-digit number beginning with 8 or 9. The rules of significant figures should be applied with common sense.

In adding numbers of different orders of magnitude, it is not the significant figures that matter but the decimal places of the digits. Thus, if we have to add three measured lengths, 176 cm, 2.846 cm, and 0.03 cm, it is clear that the digits after the decimal point in the second and third numbers do not matter, since the first number may be anywhere between 175.5 and 176.5. The last two numbers are therefore rounded off as 3 and 0, and the sum is 179. If physical quantities are to be added, they must all be expressed of course, in the same units.

In subtracting two numbers of about the same magnitude, almost all the significant figures may be lost. Thus $19177.3 - 19171.6 = 5.7$. The number of significant figures has dropped from 6 to 2. This is an important point, as some common types of statistical calculation involve such subtractions. It is therefore good practice, in a calculation where subtractions occur, to carry three or four more figures than are apparently justified by the accuracy of the data. Figures can always be dropped, but they cannot be re-inserted if they have been dropped too soon. There is nothing for it then but to start the calculation all over again.

With an automatic calculating machine it is easy, of course, to carry all the digits of which the machine is capable. This does little harm as long as the final result is properly rounded off to the degree of accuracy justified by the data.

In rounding off numbers, the last digit kept is increased by 1 if the first digit dropped is 5 or more, and is unaltered if the first digit dropped is 4 or less. If the dropped part is 5 exactly, it is usual to increase the last digit kept by 1 if it is odd and to leave it unaltered if it is even. Thus 7.1257, 3.525, and 4.735 are rounded off to three figures as 7.13, 3.52, and 4.74, respectively. This procedure insures that in many such roundings-off the numbers will be increased about half the time and decreased about half the time.

1.6 Sources of Data. Data for statistical investigation may be collected *ad hoc* (for example, by carrying out an experiment, interviewing selected individuals, or sending out questionnaires by mail), or may be obtained at second hand. The primary sources of data, those with the greatest reliability, are issued by the responsible authorities that collect them. Examples

are the publications of the Bureau of the Census or those of the Bureau of Agricultural Economics. Secondary sources of data are trade journals, newspapers, textbooks, and the like, and these should, if possible, be checked before being accepted as authoritative.

The actual collection of data is not part of the mathematical aspect of statistics, and little will be said about it here. It must never be forgotten, however, that elaborate mathematical techniques of analysis cannot compensate for bias and crudity in the original data. It is not always a simple matter to frame definitions of categories that will be free from ambiguity or questions that will elicit exactly the information required; and measurements are subject to human as well as instrumental errors. Moreover, in only a very few cases, such as a decennial census of the population, can a complete count be attempted; usually we have to be content with results from a comparatively small sample. The mathematical techniques for drawing valid conclusions about the population from such samples depend upon the assumption that the sample is random, that is, that every individual in the population has an equal chance of being included in the sample, and this is often difficult to arrange. In practice, other schemes of sampling than the purely random one are often preferred (for example, the population is sometimes divided into classes or strata, and random samples are taken from each stratum), but there must always be some element of randomness about the sampling procedure. Only thus can we be sure that the choice of individuals for the sample is independent of any personal predilections or preconceived ideas on the part of the investigator, and only then can the mathematical theory of sampling be validly applied.

1.7 Classification and Tabulation. After the data have been collected in any statistical investigation the first step has to do with introducing order in the raw material. Usually we have some hundreds of observations which have been recorded merely in the arbitrary order in which they happened to be made. But in order to analyze a set of observations so that intelligent judgments may be formed about it or so that comparisons may be made between two sets, proper classification is necessary and of prime importance.

Most people, until they have tried, imagine that to collect and arrange data in classes and in tables is a straightforward procedure involving no great technique or experience. Although much can be learned from a careful study of the illustrations and discussions that appear in the following pages and the compilations of reputable bureaus such as the census volumes, nevertheless, experience is the best teacher in effecting the most appropriate classification for any set of data.

In carrying out the process of classification, it is natural to arrange the results in tabular form, setting forth clearly and explicitly the information one wishes to present. In drawing up any table the following general rules should be observed:

(1) Every table should be self-explanatory. The title should be short, but not at the expense of clearness. It should usually tell us the "what, where, and when" of the data, in that order: thus, "Death Rates per 100,000 of Population, by Principal Causes, United States, 1949."

(2) Explanatory notes, when necessary, should be incorporated in the table, either directly under the descriptive title or directly under the body of the table.

(3) If the headings of the various columns (often called captions or box-headings) refer to descriptive categories, they should be placed, if possible, in a natural order, and usually with the most important items in the first column. A column of totals is usually placed at the right, although U. S. government publications prefer the totals on the left

The headings in the various horizontal rows constitute what is called the *stub*. The same principles apply to the stub as to the captions, the first row being generally the most important. Totals are usually (except by the U. S. government) placed at the bottom. Ordinarily a larger number of items will be placed in the stub than in the captions, as it is more convenient to run the eye down a vertical column than across a horizontal row.

(4) In tabulating long columns of figures, spaces should be left after every five or ten rows. Long unbroken columns are confusing, especially when one is comparing two numbers in a row but in widely separated columns.

(5) If the numbers tabulated have more than four or five significant figures, the digits should be grouped in threes or fours. Thus, one should write 4 685 732, not 4685732.

TABLE 1. PRINCIPAL STATISTICS OF MANUFACTURING INDUSTRIES, CANADA, 1944.
CLASSIFIED BY ORIGIN OF MATERIAL USED

<i>Origin</i>	<i>Establish- ments No.</i>	<i>Employees No.</i>	<i>Salaries and Wages \$</i>	<i>Cost of Materials \$</i>	<i>Gross Value of Products \$</i>
Farm	10,329	287,756	394,716,309	1,781,014,374	2,688,731,415
Mineral	4,479	634,542	1,208,779,764	2,258,796,792	4,708,104,244
Forest	10,347	186,680	278,171,969	495,531,476	1,082,160,284
Marine	535	9,664	10,327,695	45,906,542	68,882,879
Wild life	535	6,190	9,430,191	28,076,572	43,985,177
Mixed	2,258	98,050	128,195,442	223,007,600	481,828,520
Totals	28,483	1,222,882	2,029,621,370	4,832,333,356	9,073,692,519

Note: Mineral origin includes industries using imported iron, steel, etc.

Source: *Canada Year Book*, 1947, pp. 541-2.

(6) Double lines at the top (or at the top and bottom) may enhance the effectiveness of a table. If the table nicely fills the width of the page, no side lines should be used. In such cases the omission of the side lines will have the tendency to emphasize the other vertical lines and cause the interior columns to stand out better.

The following points are particularly important in practical work:

(7) Source of data should be included.

(8) Units of the data presented should be clear.

(9) Accuracy of transcription must not only be striven for but actually achieved. A reader who finds one error (even though this be the only one) is likely to disparage the whole table.

A specimen table* is shown in Table 1.

1.8 Frequency Distributions. From the standpoint of a mathematical analysis of statistics, the most important form of tabulation is the so-called frequency distribution. Rough data do not convey any clear idea of the

TABLE 2. GRADES OF 100 STUDENTS IN FRESHMAN MATHEMATICS

75	86	66	86	50	78	66	79	68	60
80	83	87	79	80	77	81	92	57	52
58	82	73	95	66	60	84	80	79	63
80	88	58	84	96	87	72	65	79	80
86	68	76	41	80	40	63	90	83	94
76	66	74	76	68	82	59	75	35	34
65	63	85	87	79	77	76	74	76	78
75	60	96	74	73	87	52	98	88	64
76	69	60	74	72	76	57	64	67	58
72	80	72	56	73	82	78	45	75	56

TABLE 3. FREQUENCY TABLE OF 100 GRADES

Class Limits	Tally Marks	Frequency
30-39	//	2
40-49	///	3
50-59	/// <i>///</i> /	11
60-69	<i>///</i> <i>///</i> <i>///</i> <i>///</i>	20
70-79	<i>///</i> <i>///</i> <i>///</i> <i>///</i> <i>///</i> <i>///</i> //	32
80-89	<i>///</i> <i>///</i> <i>///</i> <i>///</i> <i>///</i>	25
90-99	<i>///</i> //	7
Total		100

* Throughout the book, tables of data are presented purely as illustrations of statistical techniques and procedures. Whether the information presented in them is absolutely up-to-date is irrelevant.

TABLE 4. MONTHLY RAINFALL (INCHES) AT IOWA CITY FOR
36 CONSECUTIVE YEARS

<i>Year</i>	<i>Jan.</i>	<i>Feb.</i>	<i>Mar.</i>	<i>Apr.</i>	<i>May</i>	<i>June</i>	<i>July</i>	<i>Aug.</i>	<i>Sept.</i>	<i>Oct.</i>	<i>Nov.</i>	<i>Dec.</i>
1	2.75	0.75	1.80	1.83	2.20	7.99	0.30	2.29	1.44	2.11	1.56	0.31
2	1.49	1.30	4.41	1.11	4.46	2.80	3.01	3.45	2.33	1.63	2.93	2.72
3	1.46	1.23	3.15	4.30	9.23	8.29	6.20	2.50	1.18	1.02	1.38	2.84
4	1.18	1.75	2.82	4.37	1.79	3.01	3.56	1.64	3.07	1.98	1.75	1.52
5	1.95	1.64	2.03	2.72	3.09	2.40	0.90	2.40	4.96	2.30	1.80	0.98
6	2.37	0.64	1.25	1.66	4.26	1.10	10.10	1.77	3.43	1.38	1.78	2.84
7	0.70	1.51	0.92	5.14	4.10	1.86	7.04	2.44	1.82	2.74	1.16	0.55
8	3.66	1.30	2.07	4.60	3.11	2.38	3.83	1.85	3.54	0.33	1.98	2.48
9	4.62	1.15	3.02	2.89	4.80	3.26	2.27	2.85	2.54	4.38	1.10	0.53
10	0.59	1.82	1.43	3.23	9.49	4.50	3.78	2.39	0.93	1.66	1.15	1.93
11	0.73	2.20	3.32	3.31	4.31	2.18	5.25	6.27	4.35	3.61	1.43	0.75
12	1.07	1.97	3.62	2.36	1.54	3.33	1.29	0.66	2.56	1.78	0.79	2.34
13	1.29	0.85	1.29	1.91	3.75	7.46	6.89	10.91	5.87	3.12	2.25	2.21
14	0.67	1.03	1.86	3.11	6.90	1.95	4.76	3.45	5.38	3.60	0.97	1.27
15	1.74	0.84	2.73	5.49	2.68	2.14	2.49	3.93	3.12	1.59	0.25	1.96
16	1.22	1.90	2.28	3.36	5.37	6.68	3.59	2.62	1.54	5.36	2.92	1.04
17	2.51	1.73	2.25	1.83	2.33	3.64	1.42	5.34	0.89	1.48	3.08	1.64
18	2.12	0.22	1.59	1.58	5.47	6.04	9.21	2.98	2.85	0.86	1.07	0.53
19	0.32	2.08	2.94	2.78	7.78	2.87	5.40	7.47	1.82	1.99	1.84	0.43
20	1.97	1.09	2.00	7.21	4.40	4.58	5.75	1.88	2.43	1.59	4.88	2.52
21	1.79	0.39	0.28	2.56	3.57	0.98	2.22	4.98	3.87	0.57	0.69	0.46
22	0.87	4.82	1.30	3.02	4.74	2.98	3.70	4.27	5.07	2.78	3.01	2.29
23	0.26	1.21	2.30	3.50	2.88	2.60	3.60	3.62	2.67	3.54	1.11	0.75
24	1.19	1.42	2.69	1.83	6.91	6.28	0.39	2.97	3.19	3.66	0.46	1.02
25	1.28	0.93	2.63	2.37	4.87	5.32	1.53	2.99	7.97	1.65	0.37	1.89
26	2.15	2.42	0.92	0.65	7.65	4.33	8.11	1.80	9.31	1.84	1.80	0.80
27	3.18	0.59	5.06	1.83	5.99	3.92	1.57	2.83	3.49	3.19	1.42	1.15
28	1.09	0.19	2.19	3.43	7.33	6.49	2.84	2.79	6.23	2.28	0.30	0.57
29	1.10	1.46	0.33	3.43	6.22	8.36	4.87	6.72	2.00	2.05	2.10	1.62
30	0.08	2.63	2.65	4.28	4.49	7.07	1.03	2.67	5.10	4.01	3.84	0.61
31	0.84	1.33	4.22	4.75	3.76	2.86	2.79	2.90	1.20	0.98	1.80	2.45
32	0.35	0.49	2.46	6.20	4.44	2.46	3.59	8.61	7.83	2.47	0.74	3.19
33	1.11	1.46	2.18	3.49	5.52	0.28	6.46	1.03	2.91	1.06	5.28	0.49
34	1.09	0.67	4.83	0.86	2.63	6.21	2.37	4.01	9.27	2.35	1.13	0.73
35	1.35	0.83	2.10	1.09	1.69	8.71	3.67	5.67	2.60	1.64	0.93	1.75
36	0.29	1.04	0.99	3.07	1.06	5.61	3.63	3.14	5.59	3.90	1.00	1.66

subject matter unless they are organized and condensed in a systematic way. We therefore partition the raw data into *classes* of appropriate size, showing the corresponding frequency of variates in each class. When any set of statistics is arranged in this way it is called a frequency distribution. For example, upon a cursory examination of the raw data of Table 2 it is difficult to state any very definite conclusions as to whether these grades represent preponderantly good students or poor ones. The frequency distribution of Table 3 however, does give us more precise information. We see at a glance that there were 32 students with grades between 70 and 80, and that all but 16 had grades of 60 or above. In Table 4, the confusion of detail is still more apparent. The corresponding frequency distribution is given in Table 5.

TABLE 5. FREQUENCY TABLE OF MONTHLY RAINFALL (INCHES) AT IOWA CITY

<i>Class Interval</i>	<i>Mid-x</i>	<i>Frequency</i>
0.00- 0.49	0.245	23
0.50- 0.99	0.745	42
1.00- 1.49	1.245	58
1.50- 1.99	1.745	62
2.00- 2.49	2.245	49
2.50- 2.99	2.745	47
3.00- 3.49	3.245	32
3.50- 3.99	3.745	27
4.00- 4.49	4.245	18
4.50- 4.99	4.745	15
5.00- 5.49	5.245	14
5.50- 5.99	5.745	7
6.00- 6.49	6.245	10
6.50- 6.99	6.745	5
7.00- 7.49	7.245	6
7.50- 7.99	7.745	5
8.00- 8.49	8.245	3
8.50- 8.99	8.745	2
9.00- 9.49	9.245	5
9.50- 9.99	9.745	0
10.00-10.49	10.245	1
10.50-10.99	10.745	1
		—
Total		432

The width of a class is called the class interval, and in general the successive class intervals should be of equal width. The midvalue of such an interval is variously called the class mark, midvalue, central value. The width of a class interval is therefore seen to be the common difference between two con-

secutive class marks. It is also the difference between the lower (or upper) limit of two successive classes. Thus, in Table 5, the class interval is half an inch and the successive class marks are 0.245, 0.745, etc., inches.

For some kinds of table, equal intervals would be inconvenient. This is so in the distribution of incomes among Canadian taxpayers (Table 6), where there are so many taxpayers in the low income brackets that a comparatively fine division of incomes is feasible.

TABLE 6. INCOMES OF CANADIAN TAXPAYERS, 1929

<i>Income (dollars)</i>	<i>Frequency</i>
Under 2,000	36,857
2,000- 3,000	22,374
3,000- 4,000	19,408
4,000- 5,000	15,049
5,000- 6,000	9,529
6,000- 7,000	6,833
7,000- 8,000	3,950
8,000- 9,000	2,785
9,000-10,000	2,185
10,000-15,000	5,520
15,000-20,000	2,197
20,000-25,000	1,027
25,000-30,000	579
30,000-50,000	847
50,000 and over	523
Total	129,663

Source: *Canada Year Book*, 1930.

In the very high income groups the spread is so great that, if the same class interval were used as for the low incomes, there would be very many classes, and some of them might be empty. In such a table as this it is usual to have an *open interval* at one end or the other, that is, an interval with one of its limits indeterminate. Thus, in Table 6 the last class is "\$50,000 or more."

Another reason for unequal and open intervals, especially in government publications, is to protect anonymity. In tables dealing with output, costs, etc., for manufacturing firms, it might well happen that one or two large firms would find themselves alone in the upper classes of such tables, so that anyone with inside knowledge of the industry could identify them and so obtain confidential information. With wider classes these firms are grouped with others.

1.9 Class Intervals. Grouping variates into the most appropriate number of classes is a matter of judgment. The choice of intervals to be used in tabulating any particular set of variates depends upon the nature and characteristics of the data and the purpose for which it is to be used. For discrete

variates, the unit is a natural interval and sometimes it is satisfactory. (See Tables 10 and 11, § 2.3). However, for both discrete and continuous variates the following conditions should guide the choice:

(a) We desire to be able to treat all the values assigned to any one class, without serious error, as if they were equal to the class mark for that interval; e.g., as if all 23 items in the first class of Table 5 were exactly 0.245 inch, etc.

(b) For convenience and brevity we desire to make the interval as large as possible subject to the first condition.

These conditions will generally be fulfilled if the interval is chosen so that the whole number of classes lies between 10 and 25, depending on the total frequency. A small number of classes may "cover up" too much detail whereas a large number may reveal too much detail for one to comprehend readily (which is just the objection to the table of original data). A preliminary inspection of the data should accordingly be made and the highest and lowest values selected. Dividing the difference between these by the tentative number of classes, we have our approximate value of the interval. After a little preliminary reconnoitering an appropriate number of classes and their limits can be determined. Thus, in Table 4, the highest value noted was 10.91 and the lowest 0.08 (verify). The difference between these is 10.83, which suggests that if we took 20 classes we would have approximately a half inch as the width of a class interval. This, however, assumes we would start with 0.08 as our lower limit, which would give us awkward figures as limits. Therefore, our judgment suggests it would be better to start with 0 and continue by half-inch intervals as far as is necessary to take in the range of the given variates. We have estimated it will take approximately 20 of these; actually it turns out to be 22. This number of intervals and their width are consistent with the general conditions (a) and (b) given above.

In summary, useful rules for making a distribution are:

(1) Determine the range of the table by finding the difference between the highest value and the lowest value among the items.

(2) Determine the number of equal parts into which the range shall be divided. The size of the class interval and the number of intervals depend upon the size and nature of the distribution. (Table 3 contains rather fewer classes than is usually desirable but an interval of 10 units is quite conventional in students' grades. An interval of 5 would be used if grades of A, A-, B, B-, etc., were given instead of A, B, etc.)

(3) Arrange a sheet with three headings: class interval, tally marks, frequency.

(4) Read off the items in the raw table and for each one record a mark, as shown in Table 3.

(5) Write the sum of the marks in each row in the frequency column. The sum of the frequencies should, of course, equal the total number of observations.

Since in several calculations we assume that all the individuals in a class have the same value of the variate as the class mark, it is worth while, if there seems to be some concentration of observations around special values (particularly near the ends of the distribution), to try to arrange these concentrations to come somewhere near the middle of their class intervals. This is usually a minor consideration, however.

1.10 Class Limits and Class Boundaries. The pairs of numbers written in the column of classes of a frequency distribution, and used in tallying the original observations into their various classes, are called the *class limits*. In Table 5 the limits of the third class are 1.00 and 1.49. The measurements were, however, recorded to the nearest hundredth of an inch, so that any value between 1.485 and 1.495 would have been recorded as 1.49. Similarly, a value between 0.995 and 1.005 would have been recorded as 1.00. The third class therefore actually includes all values between 0.995 and 1.495, and these true limits are known as the *class boundaries*. Denoting the variate of Table 5 by x , the class marks by x_c (x -central), and the class boundaries by x_e (x -end), the first five classes of Table 5 are:

<i>Class Limits</i>	x_c	x_e
0 00-0 49	0.495	0.245
0 50-0 99	0.995	0 745
1 00-1 49	1.495	1.245
1 50-1 99	1.995	1.745
2 00-2 49	2.495	2.245

The intervals as determined by the *class boundaries* are adjacent, but as no recorded value can lie on a boundary there can be no ambiguity about the class to which any recorded value belongs. If the class boundaries were taken as 0.00, 0.50, 1.00, etc., as might at first appear more natural, and a recorded value happened to be 0.50, we should not know whether to put it in the first class or the second, and would have to put $\frac{1}{2}$ in each.

The distinction between class limits and class boundaries is of importance in certain calculations that are commonly made with frequency distributions. In some tables the column of classes contains only single numbers followed by dashes, as 10-, 20-, etc. These numbers are lower class limits.

1.11 Cumulative Frequencies. The frequencies so far considered may be called *absolute* frequencies, to distinguish them from *relative* frequencies (which are expressed as a fraction of the total frequency) and from *cumulative* frequencies.

Sometimes a statistical investigation is concerned with the number or percentage of observations which are "less than" or "more than" a given

value. This is frequently the case in educational tests and in wage or salary statistics. Our chief interest in such cases may be the accumulated frequency of the several class intervals up to some class boundary. Hence we are led to form a cumulative frequency table. Such a table is built up by successively adding the several (absolute) frequencies; thus: f_1 , $f_1 + f_2$, $f_1 + f_2 + f_3$, etc., as illustrated in Table 8, where the data of Table 7 are used. We shall use N to denote the sum of all the frequencies.

TABLE 7. DISTRIBUTION OF INTELLIGENCE QUOTIENTS (IQ's) OF 905 SCHOOL CHILDREN FROM 5 TO 14 YEARS OF AGE. (DERIVED FROM L. M. TERMAN, *The Measurement of Intelligence*)

<i>IQ</i>	<i>Number of Children</i>
55- 64	3
65- 74	21
75- 84	78
85- 94	182
95-104	305
105-114	209
115-124	81
125-134	21
135-144	5

TABLE 8. CUMULATIVE DISTRIBUTION OF IQ's (TABLE 7)

<i>Class Mark</i> <i>Mid-x</i>	<i>Frequency</i> <i>f</i>	<i>Upper Boundary</i> <i>End-x</i>	$F_{<}$	$\frac{F_{<}}{N}$
59.5	3 = f_1	54.5	0	0.000
69.5	21 = f_2	64.5	3 = f_1	0.003
79.5	78	74.5	24 = $f_1 + f_2$	0.027
89.5	182	84.5	102	0.113
99.5	305	94.5	284	0.314
109.5	209	104.5	589	0.651
119.5	81	114.5	798	0.882
129.5	21	124.5	879	0.971
139.5	5	134.5	900	0.994
		144.5	905 = N	1.000

The cumulative frequency corresponding to the upper boundary of any class interval is the total absolute frequency of all values less than that boundary. This is denoted by $F_{<}$ (read as "F less than"). Sometimes frequencies are cumulated from the bottom of the table, giving $F_{>}$ ("F more than"), which is the total frequency greater than a boundary value. If we divide $F_{<}$ by N , we get the *relative cumulative frequencies* of the last column of Table 8. Thus, we can readily see that about 88% of the children had IQ's less than 114.5 and only about 11% less than 84.5.

The inverse operation to cumulating the frequencies is called "differencing" and is usually denoted by Δ (delta). If S denotes any series of values, then ΔS denotes the results obtained by subtracting the first value of S from the second value, the second from the third, etc. Differencing a column of cumulative frequencies obviously gives the absolute frequencies. Differencing a column of $F_{<}/N$ values gives the f/N values.

Exercises

1. How many figures are significant in

$$(a) (132.36 - 131.64)(2.97 \times 32.2/0.0648)^{1/2}$$

$$(b) (13.189)^{1/2}(0.016524)^{1/4}/(0.03189)^2$$

if all the numbers are considered as measured values which have been rounded off?

2. In Table 3 state (a) the width of the class-interval,

(b) the class marks,

(c) the upper class boundaries.

3. Tabulate the grades of Table 2, using a class interval of 5.

4. Borrow a statistical publication such as a Company Report to Shareholders, and examine critically the tables you find in it. How far do they meet the requirements mentioned in the text?

5. (*Walker*) A study is to be made of school attendance in the 15 elementary schools of a certain city. Draw up a suitable table in which could be shown the average register, the average daily attendance and the percentage attendance, for boys and girls separately and for both, in each school

6. Consider the data of Table 9. What is the justification for the unequal class intervals?

TABLE 9. AGE DISTRIBUTION OF DEATHS OF INFANTS UNDER 1 YEAR OF AGE,
IN U.S.A. 1917 (EXCLUDING HAWAII)

<i>Age at Death</i>	<i>Frequency</i>
Under 1 day	26,665
1 day	8,364
2 days	6,344
3 to 6 days	12,375
1 week	10,911
2 weeks	7,717
3 weeks but under } 1 month	6,212
1 month	15,362
2 months	12,066
3 to 5 months	27,487
6 to 8 months	20,409
9 to 11 months	17,112
Total	171,024

N.B. Still-births are excluded.

Source: U.S. Bureau of the Census, *Mortality Statistics for 1917*.

State the class boundaries. Draw up a cumulative frequency (F_{\leq}) table. (Note that a child is reckoned as 1 month old when it has completed 1 month but has not yet completed 2 months of life, and so with the other intervals.)

7. Use Table 5 to answer (a) how often was the monthly rainfall between 2 inches and 3 inches? (put this more exactly),
(b) how often was the rainfall less than 5 inches?
(c) what is about the most common monthly rainfall?
8. Difference the last column of Table 8.

References

1. H. M. Walker, *Elementary Statistical Methods* (Henry Holt & Co., 1943), Chaps. II, III, IV.
2. S. L. Payne, *The Art of Asking Questions* (Princeton University Press, 1951).
3. F. E. Croxton, and D. J. Cowden, *Applied General Statistics* (Prentice-Hall, Inc., 1940), Chaps. I, II, III.
4. *Historical Statistics of the United States* (Bureau of the Census, 1949). A source for a number of useful tables.
5. Metropolitan Life Insurance Company, *Statistical Bulletins*. These often contain interesting frequency distributions.
6. H. M. Walker, and W. N. Durost, *Statistical Tables; Their Structure and Use* (Teacher's College, Columbia University, 1936).

CHAPTER II

GRAPHICAL REPRESENTATION

2.1 The Function Concept. Variables which are linked or related in some way are encountered in various fields of human experience. Several variables may be linked but we shall, for the present, consider the simple case where only two variables are involved. For example, the two related variables may be time and population, variate and frequency, rate of interest and accumulated principal, age and insurance premium. The primary purpose of a graph is to show diagrammatically how the values of one of two linked variables change with those of the other. One of the most useful applications of the graph occurs in connection with the representation of statistical data.

Underlying the intelligent use of graphs is the concept of *function*, which is a fundamental notion in mathematics and its applications. The student usually meets the word for the first time in algebra, when a linear or quadratic expression is spoken of as a function of x . An example is the equation

$$y = P(1 + x)^2$$

The expression on the right is the function of x (P being constant) and for convenience it is denoted by the single letter y . Here x may be an interest rate, say 0.04 or 4%, and y dollars the amount to which P dollars will accumulate in two years at 100 x % per year.

The statement that y is a function of x is written symbolically in the form

$$y = f(x)$$

This implies that a value of the function y is determined when a value is assigned to the variable x . For this reason, x is called the *independent variable* and y the *dependent variable*. In place of f other letters may be used. Thus, any one of the symbols*

$$g(x), \quad h(x), \quad F(x), \quad \phi(x)$$

and so on, denotes a function of x . The same symbol may be used to denote different functions in different problems, but different symbols are required to represent different functions in the same problem or discussion.

Examples:

$$f(x) = 5x^2 - 3x + 2$$

$$\phi(x) = Ke^{-x^2}$$

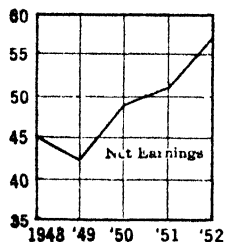
* Also, $y(x)$ is used to mean y expressed as a function of x .

Any mathematical expression involving a variable x is a function of x . However, the word is often used to designate a much more general relation. The central idea conveyed by this meaning is that of a correspondence between values of y and values of x . The following definition is the result of a development over a long period and its formulation is due to Dirichlet, a famous French mathematician (1805–59).

DEFINITION. *Let there be a set of values assumed by the independent variable x . If to each x in the set, there corresponds one or more values of y , then y is said to be a function of x in its domain.*

It should be observed that this definition is freed from any notion of the necessity of specifying the mathematical relation between x and y . A mathematical equation connecting x and y may not even exist.* A function may thus be considered as being equivalent to a table in which one may look up any x in its domain and find the corresponding y .

Many of the data in statistics come under this general definition of function. Thus, in the following table, net earning is a function of the year, whether or not there is any equation defining that functional relationship.



Year	Millions
1948	45.0
1949	43.0
1950	49.6
1951	51.5
1952	57.3

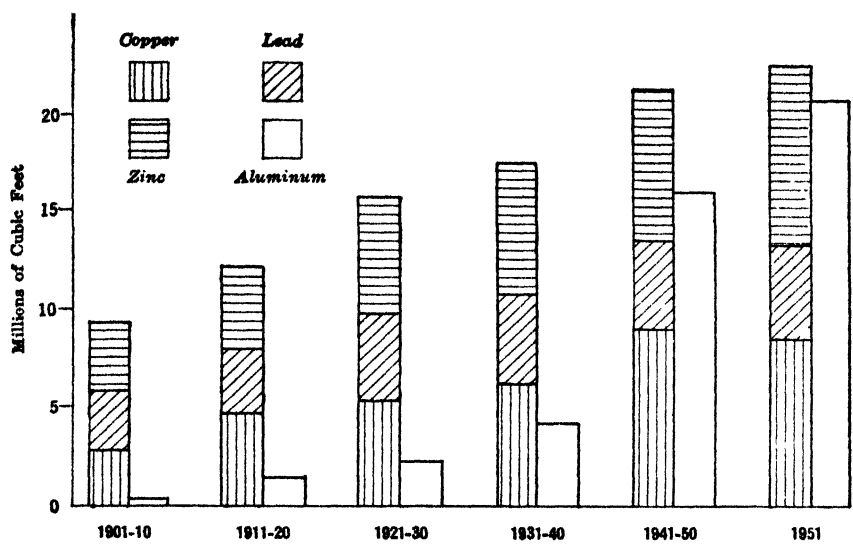
Here the function is defined only for the indicated points which correspond to the values given in the table. The straight lines are drawn to help the reader visualize the relative positions of these values and not to represent the function at intermediate points.

On the time axis a year should really be represented by an interval and not by a point, but the earnings actually spread over the whole year are here represented as a lump sum and concentrated at one point of time. The points on the broken line for intermediate times have no significance. A similar situation holds with discrete variates, such as the number of eggs laid weekly by a hen, which can obviously have only values 0, 1, 2, ...

If there is only one value of y corresponding to each value of x , y is called a *single-valued* function of x ; otherwise y is a *multiple-valued* function. In $y^2 = x$, y is a two-valued function of x , since either \sqrt{x} or $-\sqrt{x}$ will satisfy the equation. Child weight is a multiple-valued function of age, since there is a whole range of possible values of weight corresponding to a given age. The weight of a particular child is, however, a single-valued function of age.

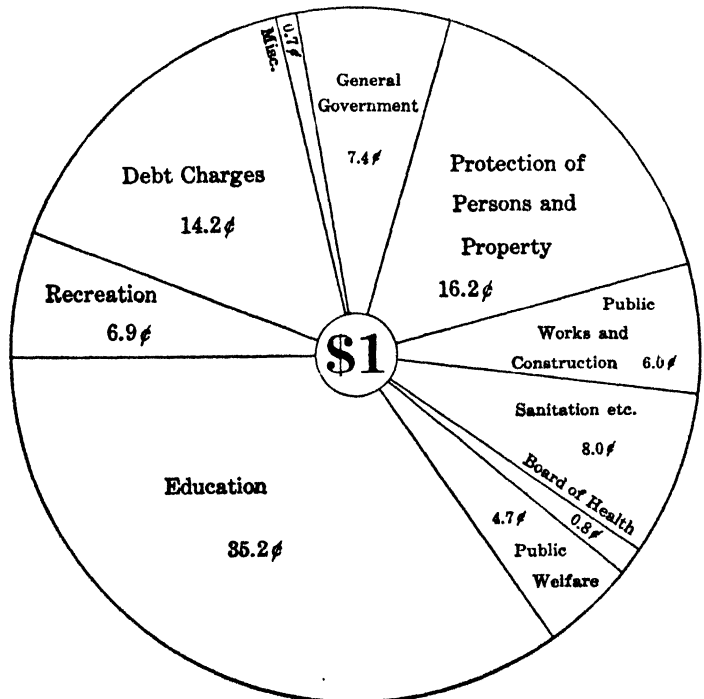
2.2 Charts. A detailed study of the technique of representing data by bar-diagrams, pie-diagrams, etc., will not be undertaken here. It is a rather specialized and nonmathematical subject, and the interested student will find ample information in references which are given at the end of the chapter.

* A classical example is the function which is defined for the infinite set of numbers from $x = 0$ to $x = 1$ to be unity for all rational numbers and zero for all irrational numbers.



Source: Annual Report of Aluminium, Ltd., 1951

FIG. 1. WORLD PRODUCTION OF NON-FERROUS METALS (EXCLUDING U.S.S.R.) — ANNUAL AVERAGES



Source: Statement by City Commissioner, 1952

FIG. 2. HOW THE CITY OF EDMONTON SPENDS ITS DOLLAR, 1951

We give one example of a bar-diagram and one of a pie-diagram to illustrate good practice (Figs. 1 and 2).

The scale of heights on a bar diagram should start from zero. Otherwise a misleading impression of the relative heights of bars may be given. The practice of replacing the bars by pictures drawn to different scales should be avoided, as it is often not clear whether the comparison is intended to be between linear dimensions, areas, or volumes. If the picture of a man is copied on double the linear scale, the area covered by the drawing is increased four times but the suggestion conveyed is that of a man with volume eight times as great. How is the picture to be understood?

2.3 Frequency Polygons. We present now a discussion of the graphs that are used in connection with frequency distributions. A distribution of values of a discrete variate may be represented graphically by plotting the points (x_1, f_1) , (x_2, f_2) , \dots (x_k, f_k) , and drawing a broken line through them. Such a graph is called a frequency polygon because it is a polygon formed by connecting the tops of a series of ordinates whose lengths are proportional to the various frequencies and whose abscissas correspond to the variate values

DICE-THROWING EXPERIMENTS (WELDON). Twelve dice were thrown 4096 times. In Table 10, only a throw of 6 was reckoned a success.

In Table 11, either 4, 5, or 6 was so reckoned.

TABLE 10

x_i	f_i
0.....	447
1.....	1145
2.....	1181
3.....	796
4.....	380
5.....	115
6.....	24
7.....	7
8.....	1
9.....	0
10.....	0
11.....	0
12.....	0
	<hr/> 4096 <hr/>

TABLE 11

x_i	f_i
0.....	0
1.....	7
2.....	60
3.....	198
4.....	430
5.....	731
6.....	948
7.....	847
8.....	536
9.....	257
10.....	71
11.....	11
12.....	0
	<hr/> 4096 <hr/>

of the distribution. Fig. 3 will serve as an illustration. For a table of discrete variates the function exists only for the given values. Likewise, its graph is discontinuous. The straight lines connecting the points serve merely to "carry the eye," thus giving a better idea of the shape and position of the distribution.

Frequency polygons are also sometimes used for distributions grouped in classes, but only when all the class intervals are equal. The frequency in any class is plotted against the class mark, and the points are joined by

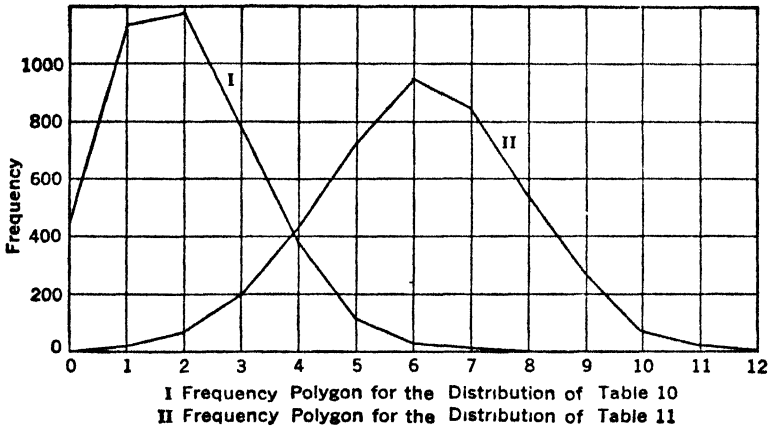


FIG. 3. FREQUENCY POLYGONS FOR DISTRIBUTION OF DISCRETE VARIATES

straight lines. The polygon should be brought down at both ends to the x -axis by joining it to the class marks of the nearest empty classes at each end of the distribution. However, it is usually preferable to use *histograms* for grouped distributions.

2.4 Histograms. A histogram is a set of rectangles with bases along the intervals between class boundaries and with areas proportional to the frequencies in the corresponding classes. If the class intervals are equal, the *heights* of the rectangles are also proportional to the frequencies, as in Fig. 4, which represents the data of Table 7.

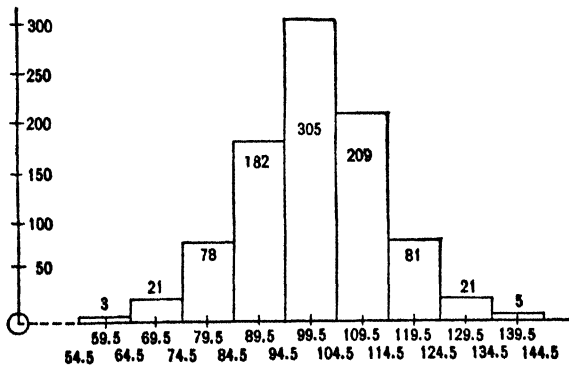


FIG. 4. HISTOGRAM FOR TABLE 7

For the data of Table 9, however, the rectangles in Fig. 5 have unequal bases and their heights are adjusted accordingly. It is *area* here, and not height, that represents frequency. To avoid excessive disproportion of heights, the data for the first month are telescoped together. A separate diagram for the deaths under 1 month might be constructed to give a truer picture.

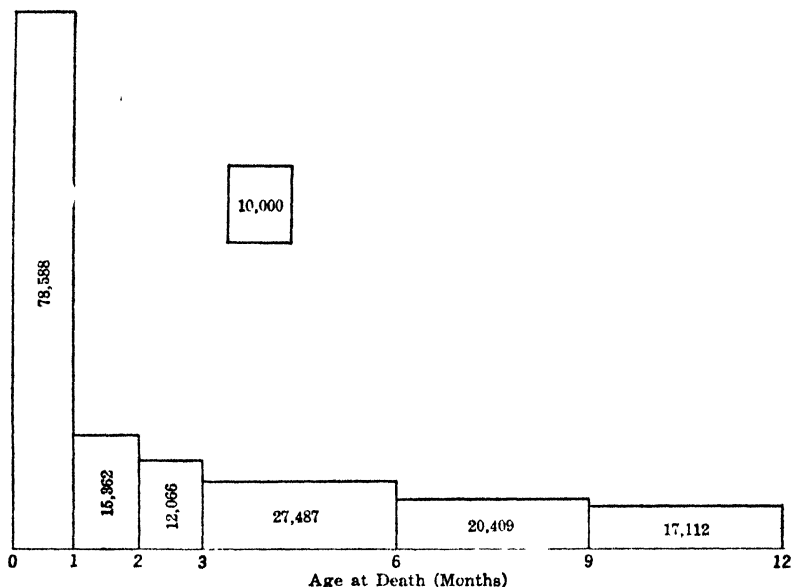


FIG. 5. AGES AT DEATH OF INFANTS UNDER ONE YEAR — U.S.A. (1917)

Note that in a histogram the rectangles are all adjacent, since the bases cover the intervals between class *boundaries*, not class limits. In a bar diagram, on the other hand, the spacing and width of the bars are arbitrary, and it is only the heights that count.

2.5 Frequency Curves. The histogram is a convenient device for representing approximately the distribution of variate values in a sample consisting of a finite number (often a few hundreds) of individuals. It is approximate because, by the method of representation by rectangles, we are assuming that within any one class the values of the variate are uniformly spread out between the class boundaries, and this is unlikely to be the case. The smaller we make the class intervals, the less likely is this assumption to be even approximately true. Moreover, with a fixed size of sample and very small class intervals there is likely to be so much fluctuation in frequency between adjacent classes, and so many of them will be empty, that the nature of the distribution is obscured. However, if we suppose that the size of the sample is increased indefinitely, so that even with very small class intervals there are many indi-

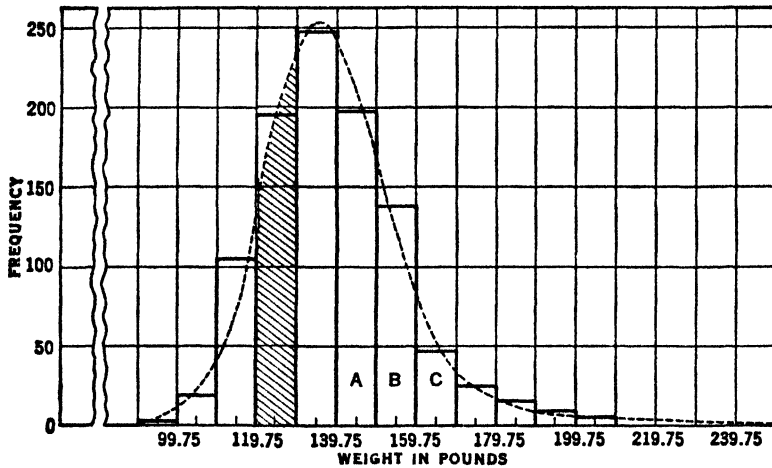
viduals in each class throughout the domain of the variate, the outline of the histogram will approximate to a smooth curve. This curve may be regarded as the *frequency curve* of the parent population from which the sample taken is a random sample. (Randomness means that any individual in the population is as likely to be picked for the sample as any other, and the population is supposed to be practically infinite in number.) In practice, frequency curves are often fitted to histograms, either by eye or more usually by calculation, using the known mathematical properties of certain curves which seem to be of about the right shape. Thus, Fig. 6 represents the data of Table 12, and

TABLE 12 — FREQUENCY DISTRIBUTION OF THE WEIGHTS OF 1000 MALE STUDENTS (ORIGINAL MEASUREMENTS MADE TO NEAREST HALF POUND)

<i>Class (Pounds)</i>	<i>Class Mark</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>
90-99.5	94.75	2	2
100-109.5	104.75	21	23
110-119.5	114.75	104	127
120-129.5	124.75	196	323
130-139.5	134.75	248	571
140-149.5	144.75	197	768
150-159.5	154.75	133	901
160-169.5	164.75	47	948
170-179.5	174.75	25	973
180-189.5	184.75	14	987
190-199.5	194.75	7	994
200-209.5	204.75	4	998
210-219.5	214.75	0	998
220-229.5	224.75	0	998
230-239.5	234.75	1	999
240-249.5	244.75	1	1000

a dotted frequency curve has been superimposed on the histogram. The total area under the curve, like the total area of the histogram, represents the total frequency (in this case, 1000). Moreover, the area between two ordinates of the frequency curve, as a fraction of the whole area, represents the *probability* that an individual selected at random from the population will have a variate value between the corresponding abscissas. In Fig. 6 the shaded area represents that fraction of the *population* which has a weight between 119.75 and 129.75 lb. The population in this example is certainly not infinite, but it may be regarded as very large. It includes all persons in the category from which the actually measured sample of 1000 may be regarded as drawn (say all white American male students attending university at the time the sample was taken). The *observed* relative frequency corre-

sponding to the same interval (represented by the area of the rectangle of the histogram on the same base, as a fraction of the total area) will differ more or less from this probability because of the inevitable variation between one sample and another, even though picked at random from the same population.



Frequency Distribution of the Weights of 1000 Male Students (Table 12)

FIG. 6

If the fit of the frequency curve to the histogram is reasonably good, these differences will not be excessive. A method of judging whether the fit is satisfactory or not will be given later after the *normal* curve has been discussed, this curve being one of the most easily fitted types of theoretical frequency curve. Many commonly occurring frequency distributions can be approximately represented by symmetrical or skew humpbacked curves with mathematically determined properties. More extreme types of distribution may be J-shaped, like that of Fig. 5, and even U-shaped distributions (with lower frequencies in the middle than at the ends) are occasionally encountered. A discussion of some theoretical curves often used in curve-fitting (Pearson curves and Gram-Charlier curves) may be found in Part Two.*

2.6 Cumulative Frequency Polygons. If the cumulative frequency $F_{<}$ is plotted against the upper class boundary (x_u) and the points are joined by straight lines, we obtain a cumulative frequency polygon. The polygon should start from zero at the lower boundary of the first interval. Fig. 7 gives the cumulative frequency polygon for the data of Table 8. Strictly speaking, $F_{<}$ for a continuous variate is defined for the end values x_u only,

* Kenney and Keeping, *Mathematics of Statistics*, Part Two, D. Van Nostrand Co., Inc., New York, 1951.

but if the assumption is made that the observations in any one class are uniformly spread out over the whole interval, the intermediate points on the polygon also represent the cumulative frequencies at the corresponding values of x . This means that we can interpolate linearly between the class boundaries, and we shall do this in the next chapter when calculating medians, quartiles, etc. The straight sides of the polygon are more than just a device to carry the eye, as in a broken-line diagram.

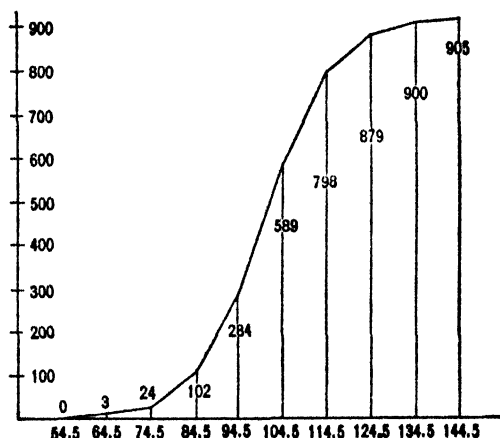


FIG. 7. CUMULATIVE FREQUENCY POLYGON

If we are dealing with a discrete variate, as in Table 11, there will be a sudden jump in the cumulative frequency at each observed value of x . The cumulative frequency polygon is stepped, as shown in Fig. 8. Corresponding to a value of x at which there is a jump, there are two values of the cumulative frequency, one $F_<$ ("less than") and the other $F_≤$ ("less than or equal to"). The difference is the frequency at this value of x . It is conventional to define the cumulative frequency for all values of x as the total frequency *up to and including* the given value, that is, to use $F_≤$. The upper end of each vertical rise on the staircase curve of Fig. 8 defines the cumulative frequency for the corresponding x . For continuous distributions it makes no difference whether we use $F_<$ or $F_≤$.

If relative frequencies instead of absolute frequencies are used the cumulative polygon rises from the value 0 at the left to the value 1 at the right. These numbers are often multiplied by 100 to give *percentage cumulative frequency polygons*.

2.7 Ogive Curves. If a large number of boys stand in a straight line, in order of height, as represented diagrammatically in Fig. 9, the line joining the tops of their heads is an approximate cumulative frequency curve, the frequency being in this case measured horizontally and the height vertically.

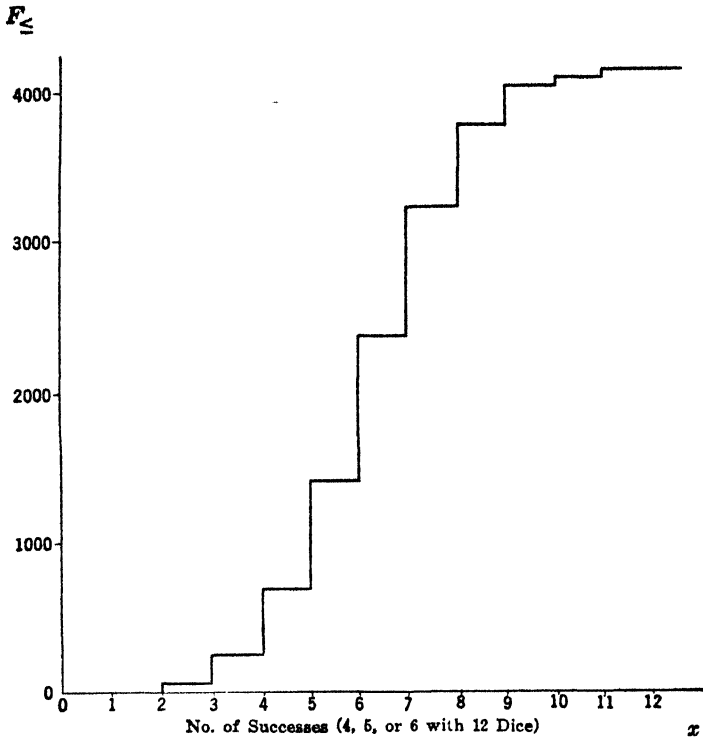


FIG. 8. CUMULATIVE FREQUENCY POLYGON

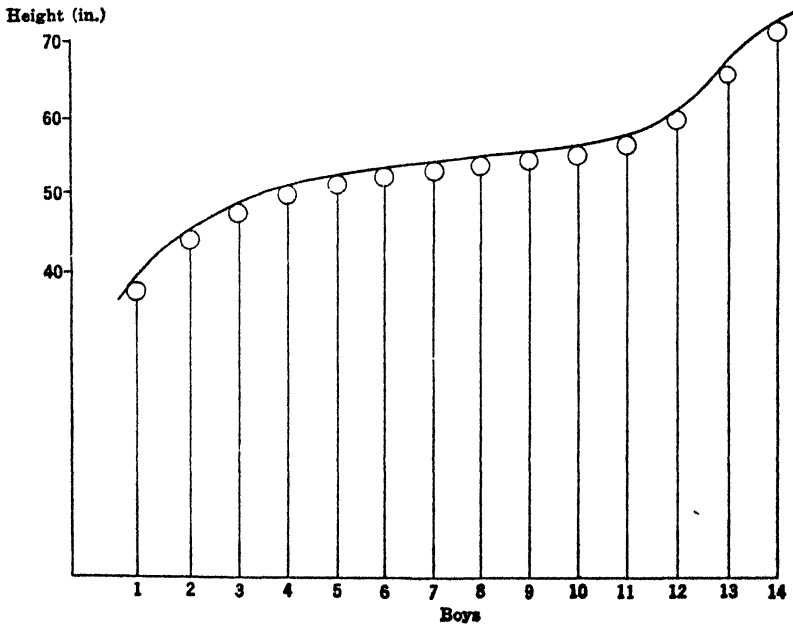


FIG. 9. Ogive

Thus, in the diagram there are 3 boys with a height less than 50 inches. From its resemblance to the shape of a moulding in architecture, known as an ogree, this curve was called an *ogive*. The name is now applied to any continuous cumulative frequency curve, including the ordinary cumulative frequency polygon. Just as the histograms and frequency polygons can often be approximately fitted by smooth curves, so the corresponding cumulative frequency polygons can be approximately fitted by smooth ogives. Indeed, in practice it is usually easier to fit an ogive than a frequency curve.

Note that the *ordinate* at any point of an ogive (plotted with the axis of x horizontal) is equal to the *area* under the corresponding frequency curve from the lowest value of x up to the value corresponding to this ordinate. Thus in Fig. 6, the area under the dotted curve up to $x = 119.75$ would be the ordinate for a smooth ogive at the same value of x . If relative instead of absolute frequencies are plotted, the ordinate gives the approximate *probability* of a value of x less than 119.75 in the parent population.

Exercises

1. If $f(x) = ke^{-x^2}$, show that $f(x) = f(-x)$.
What is the value of $f(0)$?
2. If $\phi(x) = ax^2 + bx + c$ and $\phi(x) = \phi(-x)$, show that $b = 0$.
3. If $f(x) = a^x$, show that $f(u) \times f(v) = f(u + v)$.
4. If $g(x) = \log \{(1 - x)/(1 + x)\}$, show that $g(u) + g(v) = g\{(u + v)/(1 + uv)\}$.
5. Toss four coins together and note the number of heads (x). Do this 50 times and count the number of times that $x = 0, 1, 2, 3, 4$. Construct a frequency polygon to exhibit these results.
6. Make a histogram for the data of Table 5 (§1.8). Regroup the data so that the class interval is 1 inch, and make a new histogram.
7. Construct a histogram for the data of Table 9, Ex. 6, Chap. I, covering the deaths of infants up to ages of 1 month. (Use the day as unit, and treat the month as 30 days.)
8. Draw a cumulative frequency polygon for the data of Table 9
9. Construct a cumulative frequency polygon for the data of Table 12. Draw by eye a smooth ogive. Estimate from the ogive the probability of a weight less than 130 lb in the population from which this sample was taken.

References

1. W. C. Brinton, *Graphic Methods for Presenting Facts* (Engineering Magazine Co., New York, 1914).
2. A. C. Haskell, *How to make and use Graphic Charts* (Codex Book Co., New York, 1920).
3. *Time-series Charts: A Manual of Design and Construction* (American Standards Association, New York, 1938).
4. H. Arkin, and R. R. Colton, *Graphs: how to make and use them* (Harper & Brothers, 1940).
5. F. E. Croxton, and D. J. Cowden, *Applied General Statistics* (Prentice-Hall, Inc., 1940), Chaps. IV, V, VI.
6. R. W. Burgess, *Introduction to the Mathematics of Statistics* (Houghton Mifflin Co., 1927), pp. 61-72. This gives a discussion of ogives.

CHAPTER III

THE MEDIAN AND OTHER QUANTILES

3.1 Averages. If 50 men and 50 women were selected at random from the inhabitants of any small town and their heights were measured, it is likely that, although both groups would show some variation, the men would be taller *on the average* than the women. An *average* is a value which is intended to be in some sense typical of a whole distribution. It is a more or less central value and may be regarded as a *measure of location* of the distribution on the axis of the variate x . Thus, if the two distributions (men and women separately) were plotted as histograms on the same axis they would overlap, but the one for men would be more to the right than the one for women. One of the simplest averages or measures of location is the *median*.

3.2 The Median. The median is defined as the central value of the distribution, a value such that greater and smaller values occur with equal frequency. If N values of x are arranged in order from the least to the greatest, so that

$$x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_N$$

the median is the value x_k if N is odd ($N = 2k - 1$), and is not uniquely defined if N is even ($N = 2k$) (unless $x_k = x_{k+1}$, in which case the common value is the median). If $x_k < x_{k+1}$ the median is conventionally taken half-way between, as $\frac{1}{2}(x_k + x_{k+1})$. Thus, for the numbers 5, 6, 10, 15, 18, 20, 25, the median is 15, which is x_4 . If we add another value 37, the median is between 15 and 18 and is taken to be 16.5.

For a frequency distribution of a continuous variate, grouped in classes, the

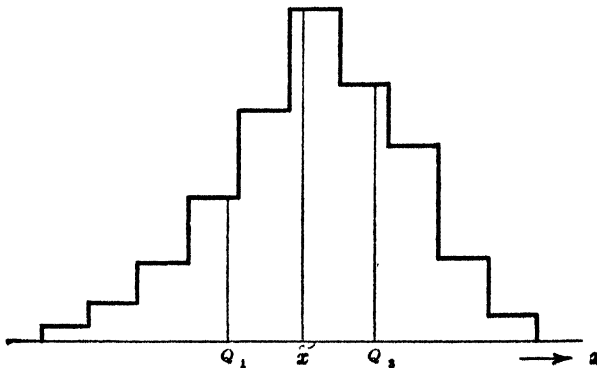


FIG. 10. MEDIAN AND QUANTILES

median is that value of x at which the ordinate divides the histogram into two parts of equal area (Fig. 10). The median is sometimes denoted by \tilde{x} (which suggests that it is a value of x) and sometimes by M_d .

As the central value in the distribution, \tilde{x} is that x for which the relative cumulative frequency is exactly 0.5. Thus, if the data of Table 3 are plotted as a cumulative frequency polygon (as in Fig. 11), the median is that value of x which corresponds to a cumulative frequency of 50 (since $N = 100$) or

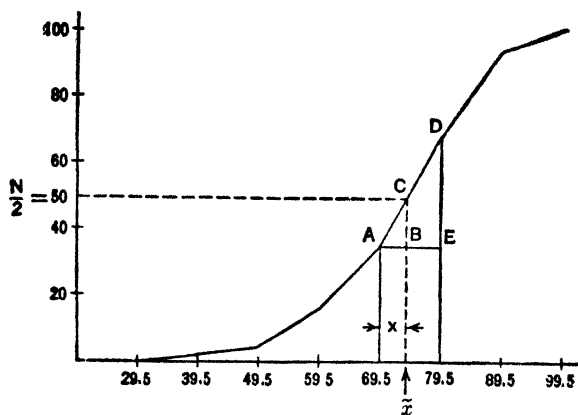


FIG. 11

a *relative* cumulative frequency of 0.5. On the assumption that, in the interval between 69.5 and 79.5, the grades are evenly distributed, the median is the abscissa of the point C. This is clearly equal to $69.5 + AB$. But, by the properties of similar triangles,

$$\frac{AB}{AE} = \frac{BC}{ED}$$

and we know that $AE = 10$, $BC = 50 - 36 = 14$, and $ED = 68 - 36 = 32$. It follows that

$$AB = 140/32 = 4.4$$

so that

$$\tilde{x} = 69.5 + 4.4 = 73.9$$

Of course, in this particular example, we have in Table 2 the original data before grouping, and thus can determine the median exactly. By arranging the grades in order, we find that the 50th and 51st are both 75, so that this is the true median. The foregoing method is useful when we are given only the grouped distribution.

Except for the purposes of illustration it is not necessary to plot the cumulative polygon in order to calculate the median. Thus for the same data

(Table 3) we need only form a column of cumulative frequencies alongside a column of upper class boundaries, and interpolate, as follows:

<i>Interval</i>	<i>f</i>	<i>End-x</i>	<i>F_{<}</i>
		29.5	0
30-39	2		
		39.5	2
40-49	3		
		49.5	5
50-59	11		
		59.5	16
60-69	20		
		69.5	36
70-79	32	$\leftarrow \tilde{x}$	$\leftarrow 50$
		79.5	68
80-89	25		
		89.5	93
90-99	7		
		99.5	100

Here, $N/2 = 50$. This value of $F_{<}$ corresponds to a value of x in the interval 69.5-79.5. Therefore the median is 69.5 plus a fraction of the distance from 69.5 to 79.5. Thus,

$$D_1 \left[\begin{array}{c} x_s \\ d_1 \left[\begin{array}{c} 69.5 \\ \tilde{x} \\ 79.5 \end{array} \right] \end{array} \right] \left| \begin{array}{c} F_{<} \\ 36 \\ 50 \\ 68 \end{array} \right. d_2 \right] D_2$$

Assuming that the items in any class interval are uniformly distributed over that interval, it follows that the partial differences are proportional to the total differences: $d_1/D_1 = d_2/D_2$. That is,

$$\frac{\tilde{x} - 69.5}{79.5 - 69.5} = \frac{50 - 36}{68 - 36}$$

whence

$$\begin{aligned} \tilde{x} &= 69.5 + 10 \left(\frac{14}{32} \right) \\ &= 69.5 + 4.4 = 73.9 \end{aligned}$$

This process is called *linear interpolation*, or *interpolation by proportional parts*, since it is equivalent to the geometrical method first described.

For a continuous variate corresponding to a random variable which is measured in certain units, the median is expressed in the same units. Thus, for a sample of heights in inches, the median is also expressed in inches.

For a *discrete variate* the median may not have much meaning. Thus in Table 11 we have $N/2 = 2048$, and the relevant portion of the cumulative frequency table is

x	$F_<$	$F\leq$
5	695	1426
6	1426	2374
7	2374	3221

The 2048th value of x is clearly 6, which is the median according to the first definition of this section, but there are 948 observed values all equal to 6. There are, however, only 1426 values less than 6 and 1722 values greater than 6.

3.3 Quartiles. Just as the ordinate at the median of a grouped distribution divides the histogram into two parts of equal area, so the ordinates at the quartiles Q_1 and Q_3 cut off one-quarter of the area at each end (Fig. 10).

The first quartile, denoted by Q_1 , is that value of x for which $F_< = N/4$. That is, one-fourth of all the variates in the distribution are smaller in value than Q_1 and three-fourths of them are larger than Q_1 . The second quartile Q_2 is that value of x for which $F_<$ is $N/2$ and is therefore the median. The third quartile, denoted by Q_3 , is that value of x for which $F_< = 3N/4$. Hence 50% of the total frequency is included between Q_1 and Q_3 . On the relative cumulative frequency polygon for a continuous variate, the ordinates at Q_1 , Q_2 , Q_3 are 0.25, 0.50, and 0.75, respectively.

Like the median, the quartiles are calculated by interpolation in the cumulative frequency table, as illustrated by the following example.

Example. (a) Find the median and the quartiles for the distribution of IQ's in Table 7 (§1.11). (b) Illustrate the measures found in (a) by means of a $F_<$ graph.

x_i	$F_<$
54.5	0
64.5	3
74.5	24
84.5	102
$\leftarrow Q_1$	
94.5	284
$\leftarrow Q_2$	
104.5	589
$\leftarrow Q_3$	
114.5	798
124.5	879
134.5	900
144.5	$N = 905$

Solution:

$$N/4 = 226.25, \quad N/2 = 452.5, \quad 3N/4 = 678.75$$

$$\frac{Q_1 - 84.5}{10} = \frac{226.25 - 102}{284 - 102}, \quad Q_1 = 91.33$$

$$\frac{Q_2 - 94.5}{10} = \frac{452.5 - 284}{589 - 284}, \quad Q_2 = 100.02$$

$$\frac{Q_3 - 104.5}{10} = \frac{678.75 - 589}{798 - 589}, \quad Q_3 = 103.79$$

$$Q = \frac{Q_3 - Q_1}{2} = 8.73$$

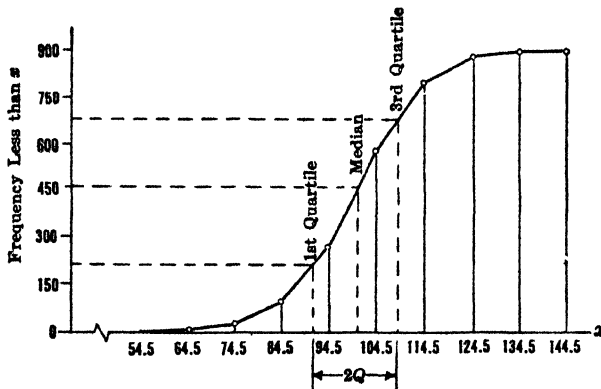


FIG. 12

Fig. 12 explains graphically the measures obtained by interpolation from a $F_<$ table. For convenience in drawing the figure, the quartile labels are put on vertical lines. But one should remember that the quartiles are values of x and that it is the horizontal distances of the lines from the y -axis that represent these measures.

3.4 The Quartile Deviation. Half the distance between Q_1 and Q_3 is called the *semi-interquartile range* or the *quartile deviation* and may be denoted by Q . Thus,

$$Q = (Q_3 - Q_1)/2$$

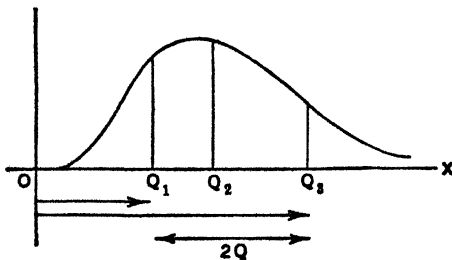


FIG. 13

The median is not necessarily midway between Q_1 and Q_3 (see Fig. 13), although this will be so for a symmetrical distribution.

For a variable which is measured in certain units (such as a height in inches), the quartile deviation is expressed in the same units.

The purpose of calculating the quartile deviation is to have a measure of the *dispersion* of the distribution, that is, of the way in which it is spread out along the x -axis. Some distributions show a marked tendency to bunch around a central value, whereas others are more nearly uniformly spread over a finite interval. Fig. 14 shows that two distributions may have the same

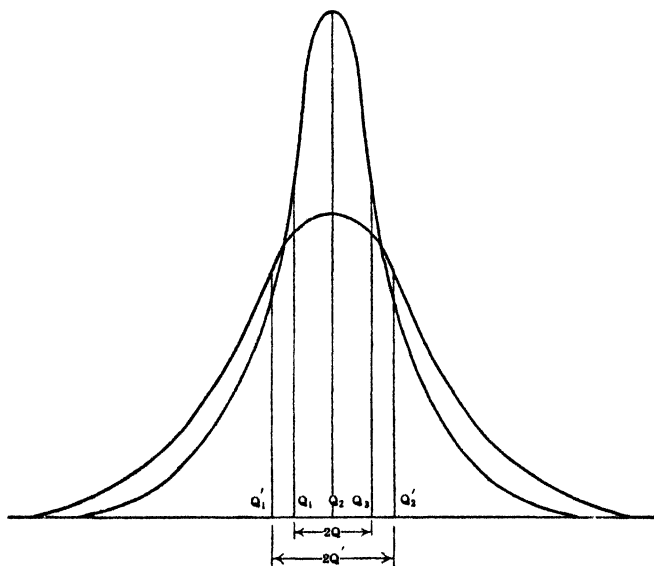


FIG. 14. DISTRIBUTION WITH DIFFERENT DISPERSIONS

median and total frequency, yet differ considerably in dispersion. The different quartile deviations for these two indicate the difference in dispersion.

3.5 Quantiles. We can calculate many other quantities* similar to the quartiles, but corresponding to different fractions of the total frequency. These statistics are collectively called *quantiles*, or sometimes *fractiles*. For example, in educational and psychological work *deciles* are often calculated. The deciles D_1 to D_9 are values of the variate x (e.g., scores on an intelligence test) such that one-tenth of all the values obtained lie below D_1 , one-tenth between D_1 and D_2 , and so on. Thus we can say that any person who has taken the test is in the top (or some other) tenth of the whole group tested, and thus form a readier appraisal of his performance. With a large sample it may be useful to calculate *percentiles*. The k th percentile P_k is that value of x , say x_k , which corresponds to a cumulative frequency of $Nk/100$. The

* A quantity such as a median, quartile deviation, etc., which is calculated from the observed data, is often called a "statistic." This is a useful technical sense of the term, but one which is quite distinct from the various meanings of "statistics" discussed in §0.1. See also §7.9.

50th percentile is obviously the median, the 25th percentile is Q_1 , the 20th percentile is D_2 , and so on. The calculation of any percentile by linear interpolation is precisely similar to that of the quartiles.

One objection to the quartile deviation as a measure of dispersion is that it is not very sensitive to differences affecting mainly the ends of the distribution. All the x values lying beyond the class interval containing Q_3 , for instance, might be increased arbitrarily without changing the position of Q_3 at all. For this reason it is often felt that the distance from P_7 to P_{93} is a better measure of dispersion, since comparatively few observations lie outside these limits. Of course, the whole range from the least to the greatest value of x could be used, but it suffers from the defect of being unduly sensitive to sampling variation. That is, the range depends a great deal on the chance presence or absence in the sample of a few extreme values of the variate.

3.6 Percentile Ranks. The percentile rank of P_k is k . Thus, instead of saying that the 20th percentile is 57, we could say that the percentile rank of 57 is 20. Both statements mean that in the particular sample studied 20% of the individuals had a value of the variate x less than 57. If $F_<$ is the cumulative frequency corresponding to the k th percentile, then

$$F_</math>$$

The relation between percentiles and percentile ranks is that between abscissas and ordinates of a percentage cumulative frequency polygon, as illustrated in Fig. 15.

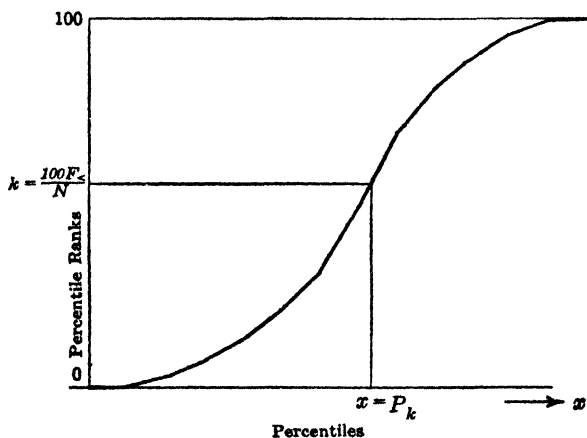


FIG. 15. PERCENTILES AND PERCENTILE RANKS

The calculation of a percentile rank is similar to that of a percentile, but the interpolation is in the column of cumulative frequency rather than in the column of upper class boundaries. Suppose we want to find the percentile rank of 110 in the data of the example in §3.3; that is, we want the percentage

of the school children studied with an IQ below 110. Now, 110 lies in the interval between 104.5 and 114.5, and the values of $F_{<}$ for the ends of this interval are 589 and 798. Hence the value of $F_{<}$ corresponding to 110 is given by

$$\frac{F_{<} - 589}{798 - 589} = \frac{110 - 104.5}{114.5 - 104.5} = 0.55$$

so that $F_{<} = 704$. This, as a percentage of 905, is 77.8, which is the required percentile rank. It would usually be rounded off as 78.

In the field of education, percentile ranks are often referred to as *grades*.

3.7 Approximate Characterization of a Distribution by Quantiles. If in a large sample we know the median, the quartiles, and the two extreme values, we can form a pretty good idea of the whole distribution by plotting these five points and sketching in a percentage cumulative frequency curve joining them. As shown in Fig. 16, these five points are equally spaced along the

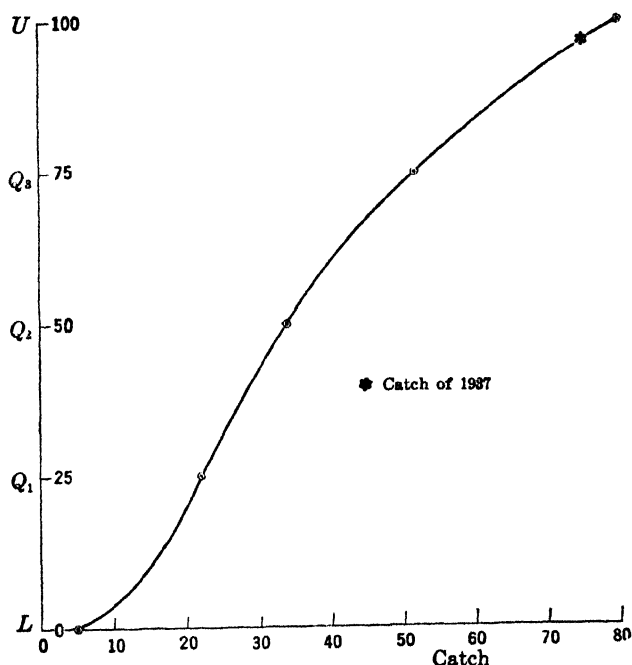


FIG. 16. CATCH OF SOLES, ABERDEEN, 1912-1937

vertical axis. The data of the figure refer to the total catch of soles* at Aberdeen, Scotland, over a period of 26 years (units not stated).

The catch in 1937, marked by an asterisk, is seen to correspond to a per-

* D'Arcy W. Thompson, *Nature*, 144 1939, p. 445.

centile rank of 96. On the basis of the 26 years studied we should expect the annual catch to exceed this value (75) only about once in 25 years (4%).

If a set of data is too scanty to permit forming a frequency distribution, an approximate ogive can be drawn by arranging the variate values in order, giving them cumulative frequencies of $\frac{1}{2}$, $1\frac{1}{2}$, $2\frac{1}{2}$, and so on, and plotting these frequencies (expressed as percentages) against the corresponding x . Thus, Table 13 gives the maximum annual flow* in the North Saskatchewan River

TABLE 13. MAXIMUM ANNUAL FLOW OF NORTH SASKATCHEWAN RIVER AT EDMONTON
(Percentage of Mean)

x	$F <$	$\%F <$	x	$F <$	$\%F <$
66.1	0.5	2.8	102.5	9.5	52.8
67.0	1.5	8.3	105.0	10.5	58.3
73.2	2.5	13.9	113.7	11.5	63.9
75.3	3.5	19.4	116.8	12.5	69.4
83.8	4.5	25.0	117.8	13.5	75.0
85.6	5.5	30.6	120.6	14.5	80.6
88.1	6.5	36.1	123.4	15.5	86.1
98.4	7.5	41.7	124.0	16.5	91.7
99.2	8.5	47.2	138.8	17.5	97.2

for each of 18 years, placed in increasing order. The value $x = 83.8$ is given the cumulative frequency of 4.5 because there are 4 values less than, say,

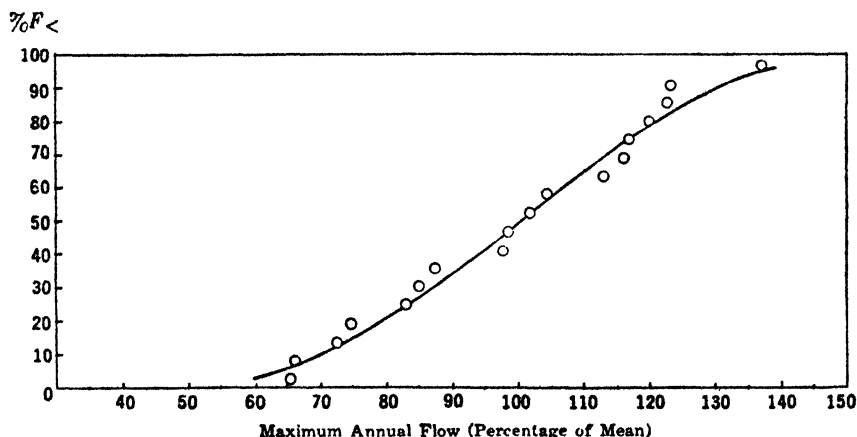


FIG. 17. FLOW OF NORTH SASKATCHEWAN RIVER, EDMONTON

* The flow is expressed as a percentage of the mean (see Chapter IV). The units do not matter for the present purpose.

83.7 and 5 values less than 83.9. The value 83.8 is, so to say, split in two and reckoned half with the lower values and half with the higher values. The plotted points do not lie on a smooth curve, but an approximate ogive may be sketched so as to lie fairly evenly between them (Fig. 17). This ogive may be used to estimate the chances of a flow of any given size.

Exercises

1. Calculate the median and quartiles, and the quartile deviation, for the distribution of Table 12 (§2.5).

2. Calculate the median income in Table 6 (§1.8). Note that the median can be found even though there are open classes at the ends of the distribution.

3. Use the cumulative frequency polygon of Exercise 8 (Ch. II) to find approximate values of the median and quartiles of age at death of infants dying under 1 year. Note that this distribution is very asymmetrical (skew) and that consequently $Q_3 - Q_2$ is very different from $Q_2 - Q_1$.

4. Explain why the median is found from interpolating in the end- x column (x_e) and not in the mid- x column (x_c).

5. Criticize the following "definitions":

$$Q_1 = N/4, \quad Q_2 = N/2, \quad Q_3 = 3N/4.$$

6. Compute the 3rd decile and the 65th percentile for the distribution of Table 12 (§2.5).

7. Find the percentile ranks of 120 lb and 200 lb in the distribution of Table 12. Interpret your answers.

41727



CHAPTER IV

THE ARITHMETIC MEAN AND OTHER AVERAGES

4.1 Various Averages. As already pointed out in §3.1, an average of a distribution is a more or less typical value of the variable, used to characterize the location of the distribution as a whole. It is in some sense a *central value* of the distribution, although it need not actually be in the domain of the variable. Thus the average number of children in a group of families may be 2.7, although obviously the number of children in a family cannot be fractional.

One average, the *median*, has already been considered, but there are several others in common use. The most important by far is the *arithmetic mean*, and a considerable portion of this chapter will be devoted to it. We shall also describe briefly the *mode*, the *geometric mean*, the *harmonic mean*, and the *root mean square*. As a preliminary we introduce some notation which will be useful.

4.2 Notation for Sums and Products. If x denotes a variable, then x_1, x_2, \dots, x_N , are symbols for the values which x may take. When we are concerned with a sum like the following,

$$x_1 + x_2 + x_3 + x_4 + \dots + x + \dots + x_N$$

it is customary to designate it by placing the Greek capital letter Σ (sigma) before the general term, thus*

$$\sum_{i=1}^N x_i = x_1 + x_2 + \dots + x_i + \dots + x_N$$

The symbol Σ is a sort of mathematical verb, and the notation written above and below it may be called adverbs. Mathematicians call Σ an *operator* and speak of the "adverbs" as *limits*. When $\sum_{i=1}^N$ is placed before any quantity, it means, "add up all quantities like \dots which are formed by giving i the values of every positive integer from $i = 1$ to $i = N$, inclusive." Thus if x_i stands for a value of the variate in Table 2, x_1 refers to the first value 75, x_2 refers to the second value 80, etc., and x_N refers to the last value 56. Here $N = 100$. Hence the compact notation $\sum_{i=1}^{100} x_i$ denotes the sum of all the values in Table 2. The symbol $\sum_{i=1}^N x_i$ is read, "the summation of x -sub- i , i varying (or running) from one to N ." The subscript i is called the *index of summa-*

* The symbols Sx and $S(x)$ may also be used instead of Σx .

tion. Any letter may be used as an index but it is conventional to use i or j . Also the upper limit may be denoted by any letter, but we shall use N to denote the total number of observed values (some of which may be alike) in a set.

If a variable x is to take on the particular values, 1, 2, 3, etc., instead of the general values x_1, x_2, x_3 , etc., then x itself becomes the index of summation and we write $x = 1$ underneath \sum . Thus

$$\sum_{x=1}^N x = 1 + 2 + 3 + \cdots + N$$

Frequently the index of summation is understood from the context and the notation at the top and bottom of \sum may be omitted if no ambiguity results.

Illustrations:

1. $\sum_{i=1}^N 3x_i = 3x_1 + 3x_2 + \cdots + 3x_N$
 $= 3(x_1 + x_2 + \cdots + x_N).$
2. $\sum_{i=1}^5 (x_i + c) = (x_1 + c) + (x_2 + c) + (x_3 + c)$
 $+ (x_4 + c) + (x_5 + c)$
 $= (x_1 + x_2 + x_3 + x_4 + x_5) + 5c.$
3. $\sum_{i=1}^4 x_i f_i = x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4.$
4. $\sum_{j=1}^4 x_j y_j = x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4.$
5. $\sum_{u=1}^N u^2 = 1^2 + 2^2 + 3^2 + \cdots + N^2.$

The following theorems may be proved very simply by writing out the summations in full:

Theorem 1. $\sum_{i=1}^N (x_i + y_i - z_i) = \sum_{i=1}^N x_i + \sum_{i=1}^N y_i - \sum_{i=1}^N z_i$

Theorem 2. *If c is a constant,*

$$\sum_{i=1}^N c x_i = c \sum_{i=1}^N x_i$$

Theorem 3. $\sum_{i=1}^N c = c + c + \cdots + c = Nc$

The next two theorems are concerned with the summing of positive integers and their squares.

Theorem 4.
$$\sum_{x=1}^N x = N(N+1)/2$$

This follows from the fact that the sum is an arithmetic progression with first term 1 and common difference 1.

Theorem 5.
$$\sum_{x=1}^N x^2 = N(N+1)(2N+1)/6$$

Proof: Let us take the identity $x^3 - (x-1)^3 = 3x^2 - 3x + 1$, and sum each side for $x = 1$ to N . Thus,

$$\sum_{x=1}^N [x^3 - (x-1)^3] = \sum_{x=1}^N [3x^2 - 3x + 1]$$

Applying Theorems 1-3 to the right member we have

$$\sum_{x=1}^N [x^3 - (x-1)^3] = 3 \sum_{x=1}^N x^2 - 3 \sum_{x=1}^N x + N$$

Performing the indicated sum in the left member, we have

$$\left. \begin{array}{l} 1^3 - 0^3 \\ 2^3 - 1^3 \\ 3^3 - 2^3 \\ \vdots \\ N^3 - (N-1)^3 \end{array} \right\} \text{whose sum is } N^3$$

Therefore

$$N^3 = 3 \sum x^2 - 3 \sum x + N$$

Hence, using Theorem 4 and simplifying,

$$\begin{aligned} \sum_{x=1}^N x^2 &= \frac{2N^3 + 3N(N+1) - 2N}{6} \\ &= \frac{N(N+1)(2N+1)}{6} \end{aligned}$$

If we wish to denote the *product*, instead of the sum, of the values x_1, x_2, \dots, x_N , we use the Greek capital letter Π (pi), thus

$$\prod_{i=1}^N x_i = x_1 x_2 \cdots x_N$$

Some simple theorems follow immediately from this definition; for example:

Theorem 6.
$$\prod_{i=1}^N (x_i y_i) = \prod_{i=1}^N x_i \prod_{i=1}^N y_i$$

Theorem 7.
$$\prod_{i=1}^N (cx_i) = c^N \prod_{i=1}^N x_i$$

Theorem 8.
$$\prod_{x=1}^N x = 1 \cdot 2 \cdot 3 \cdots N \text{ (which is called “} N \text{ factorial”).}$$

The product notation is not used as frequently in elementary work as the sum notation.

4.3 Arithmetic Mean. The arithmetic mean of the values x_1, x_2, \dots, x_N is their sum divided by N . If we denote the arithmetic mean by \bar{x} (read as “ x -bar”),

$$(4.1) \quad \bar{x} = (x_1 + x_2 + \cdots + x_N)/N = \frac{1}{N} \sum_{i=1}^N x_i$$

Thus for the set of grades in Table 2, we find

$$\bar{x} = 7266/100 = 72.66$$

Computing the mean* strictly according to definition (4.1) may be called the serial method to distinguish it from other methods which will be presented. This definition is used when N is so small that a grouping of the values into a frequency distribution is not desirable.

A considerable amount of arithmetic may often be saved by mentally subtracting a suitable number from all the values of x before adding them. The same number is added to the mean after it has been computed. Thus, if the values are 171, 173, 175, 181, 189, 196, 197, 200, we subtract 170 from each and find the mean of the remainders 1, 3, 5, 11, 19, 26, 27, 30. This is $122/8 = 15.25$. The mean of the original numbers is therefore 185.25. The validity of this procedure follows from the relation

$$\frac{1}{N} \sum_i (x_i - c) = \frac{1}{N} \sum_i x_i - \frac{1}{N} Nc = \bar{x} - c$$

4.4 Weighted Arithmetic Mean. It will be noticed that several of the grades given in Table 2 are alike. For example, 80 occurs seven times. It should be evident that the same result would be found for the mean if, instead of summing the individual values, each value were first multiplied by the frequency with which it occurs and all such products were then added. In general, if the values x_1, x_2, \dots, x_k occur with corresponding frequencies f_1, f_2, \dots, f_k , respectively, where $f_1 + f_2 + \cdots + f_k = N$, it follows that

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k}$$

or, in shorter notation,

$$(4.2) \quad \bar{x} = \frac{1}{N} \sum_{i=1}^k f_i x_i, \text{ where } N = \sum_{i=1}^k f_i$$

* When there is no ambiguity, the arithmetic mean is often referred to as the mean.

Here each value x_i is said to be *weighted*, the weight being the corresponding frequency f_i , and the arithmetic mean so obtained is called a *weighted arithmetic mean*. The term originated in experimental science where readings are sometimes "*weighted*" according to their estimated reliability. The mean of three independent measurements of a quantity may be taken to be three times as reliable as a single measurement and given a weight 3, compared with weight 1 for a single reading. In this case also, the weight is essentially a frequency.

The ordinary arithmetic mean may be regarded as a weighted mean in which all the weights are equal. If the x 's are added individually, the f 's become unity, and equation (4.2) reduces to (4.1). The student should notice that, for the same data, $\sum_1^k f_i x_i$ is numerically equal to $\sum_1^N x_i$. He should also observe that N refers to the total number of values in the set (some of which may be alike), whereas k refers to the number of *different* values of x in the set and hence to the number of products of the form $x_i f_i$, where f_i is the number of times x_i occurs. In the following example, $N = 8$ and $k = 4$.

Example. For the values 6, 8, 7, 6, 5, 7, 6, 5,

$$\sum_{i=1}^8 x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 6 + 8 + 7 + 6 + 5 + 7 + 6 + 5 = 50$$

$$\sum_{i=1}^4 f_i x_i = f_1 x_1 + f_2 x_2 + f_3 x_3 + f_4 x_4 = 2 \cdot 5 + 3 \cdot 6 + 2 \cdot 7 + 1 \cdot 8 = 50. \quad \sum_{i=1}^4 f_i = 8$$

By either method, $\bar{x} = 50/8 = 6.25$.

The general formula for the weighted arithmetic mean is $\bar{x} = \sum w_i x_i / \sum w_i$, where the weights w_i need not be frequencies.

4.5 Arithmetic Mean of a Grouped Variate. For the purpose of calculating the arithmetic mean for a continuous variate grouped in classes, we assume that all the items falling in the same class interval have the same value of the variate, namely, the class mark for that interval. This is not the assumption we make in calculating the median, that is, that the values are uniformly spread out over the interval, but as far as the calculation of the arithmetic mean is concerned the two assumptions are equivalent.* Our mean is therefore a weighted mean, the x_i being the class marks and the f_i the class frequencies. Table 14 illustrates the calculation for the data of Table 3 (§1.8).

Since in Table 2 we have the original ungrouped data, we can calculate the arithmetic mean exactly. The result is 72.66 and shows that our assumption

* On the assumption of uniform spread, any value in the upper half-interval is matched by a corresponding value in the lower half-interval at the same distance from the class mark. The sum of these two is just twice the class mark, and therefore is the same as the sum of two values each equal to the class mark.

TABLE 14

<i>Class Limits</i>	<i>Class Mark x_c</i>	<i>Frequency f</i>	<i>Product fx_c</i>
30-39	34.5	2	69.0
40-49	44.5	3	133.5
50-59	54.5	11	599.5
60-69	64.5	20	1290.0
70-79	74.5	32	2384.0
80-89	84.5	25	2112.5
90-99	94.5	7	661.5
Totals		$\sum f = 100$	$\sum fx_c = 7250.0$

$$\bar{x} = \frac{7250}{100} = 72.50$$

in this example causes little error. With fairly large samples (at least two or three hundred), for which the class frequencies tail off gradually at both ends of the distribution, the "grouping error" of the mean caused by making this assumption is hardly ever serious, unless the class intervals are quite broad.

The calculation of the mean may be considerably simplified arithmetically by a suitable change of the variate values. This change is equivalent geometrically to shifting the origin of reference and at the same time altering the scale.

When a frequency distribution is represented by a graph, we have seen in Chapter II that the variate values are used as abscissas or measurements along the x -axis. The mean is therefore the point on the x -axis whose coordinates are $(\bar{x}, 0)$. Its position may be emphasized by drawing a vertical line through this point, but it is the horizontal distance of the point from the origin, and not the vertical line, which represents graphically the mean.

If new axes $x'y'$, are taken parallel to the old axes, xy , with positive directions preserved, the axes are said to be *translated* from one position to the other. A translation of axes corresponds to a transformation of coordinates. Thus if we let

$$x' = x - x_0, \quad y' = y - y_0$$

the origin is translated to the point (x_0, y_0) . Since our variate is denoted by x we are concerned here only with the transformation $x' = x - x_0$ which translates the origin to the point $(x_0, 0)$. The new values x' are often called *deviations*. The units of measurement remain unchanged. Obviously, any values that are larger than x_0 will be positive in terms of x' and any values smaller than x_0 will be negative in terms of x' .

We can now alter the scale by letting c units of the variate x' ($= x - x_0$) equal 1 unit of a new variate u . The number expressing an observation in terms of u will therefore be $1/c$ times as large as it would be in terms of x' , so that*, if $c \neq 0$,

$$(4.3) \quad u = (x - x_0)/c$$

The relation between \bar{x} and \bar{u} is expressed by

Theorem 9.

$$(4.4) \quad \bar{x} = c\bar{u} + x_0$$

Proof: By (4.3), $x = cu + x_0$. Substituting in (4.2) we have

$$\bar{x} = \frac{1}{N} \sum_{i=1}^k f_i (cu_i + x_0) = \frac{c}{N} \sum_i f_i u_i + \frac{x_0}{N} \sum_i f_i$$

The first term is $c\bar{u}$, by definition, and the second is x_0 , since $\sum_i f_i = N$. The theorem follows.

Returning to Table 14, we see that if we choose $x_0 = 64.5$ (the class mark of the middle class) and choose $c = 10$ (the class interval) the numbers $u = (x - 64.5)/10$ become simple integers arranged in order. The products fu are easily multiplied mentally, and then \bar{x} is given by (4.4). The procedure is illustrated in Table 15, which shows the so-called "fully coded" method of calculating \bar{x} . (The variate is *coded* to a new scale.) It is simply

TABLE 15. MEAN OF 100 GRADES, WITH CLASS INTERVAL AS UNIT

x	u	f	fu
34.5	-3	2	-6
44.5	-2	3	-6
54.5	-1	11	-11
64.5	0	20	0
74.5	1	32	32
84.5	2	25	50
94.5	3	7	21
		<hr/> 100	<hr/> 80

$$\bar{u} = 80/100 = 0.8$$

$$\bar{x} = 64.5 + 10\bar{u} = 72.5$$

a device to save arithmetic. Note that any of the class marks may be chosen as the value of x_0 , but it is convenient to choose one in the region of the higher

* This is the same kind of relation as that between the centigrade and Fahrenheit scales of temperature, namely, $C = (\frac{5}{9})(F - 32)$. Here the scale-factor c is $\frac{9}{5}$.

frequencies. This insures that the larger numbers in the f column will be multiplied by numerically small values of u .

If the class intervals are unequal, the full benefit of coding cannot be obtained, but a suitable choice of x_0 and c will usually considerably simplify the numbers which have to be multiplied. Thus in Table 16 (fictitious data), we may take $x_0 = 74.5$, $c = 5$, giving the u values shown in the fourth column.

TABLE 16. CODED CALCULATION OF MEAN WITH UNEQUAL INTERVALS

Class Limits	f	x_c	u	fu
10- 19	1	14.5	-12	-12
20- 29	5	24.5	-10	-50
30- 39	12	34.5	-8	-96
40- 49	25	44.5	-6	-150
50- 99	47	74.5	0	0
100-199	55	149.5	15	825
200-499	46	349.5	55	2530
500-999	9	749.5	135	1215
	<hr/> 200			<hr/> 4262

$$\bar{u} = 21.31$$

$$\bar{x} = 74.5 + 106.55 = 181$$

4.6 Mean of Means. One of the great advantages of the mean as an average is its amenability to mathematical treatment. For example, we shall prove a theorem connecting the mean of two sets of values, when combined into a single set, with the means of the two sets taken separately. In order to be able to generalize to more numerous subsets, it will be convenient to extend the subscript notation, and use x_1, x_2, \dots to mean different *variables* and not different values of a single variate. Then we can use a second subscript for the values. If the variate x_1 has n_1 observed values, we can denote these by

$$x_{11}, x_{12}, \dots, x_{1n_1}$$

and similarly the n_2 values of x_2 are denoted by

$$x_{21}, x_{22}, \dots, x_{2n_2}$$

(x_{11} is read "x one one" and not "x eleven", etc.)

The mean of the first set is

$$(4.5) \quad \bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$$

and the mean of the second set is

$$(4.6) \quad \bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j}$$

Theorem 10. *The mean of the combined sets, consisting of n_1 values of x_1 and n_2 values of x_2 , is*

$$(4.7) \quad \bar{x} = (n_1\bar{x}_1 + n_2\bar{x}_2)/N, \quad N = n_1 + n_2$$

Proof: From (4.5) and (4.6)

$$n_1\bar{x}_1 + n_2\bar{x}_2 = \sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2j}$$

If we denote "either x_1 or x_2 " by x , x can take on $n_1 + n_2$ values, namely, the n_1 values of x_1 and the n_2 values of x_2 , and the sum of these may be denoted by $\sum_{k=1}^{n_1+n_2} x_k$, the values from $k = 1$ to n_1 being values of x_1 and the rest being values of x_2 . If $n_1 + n_2 = N$, the combined mean is

$$\bar{x} = \frac{1}{N} \sum_{k=1}^{n_1+n_2} x_k = \frac{1}{N} \left[\sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2j} \right] = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

This result may be generalized as:

Theorem 11. *The mean of a set of N values which is composed of k subsets, the frequency in the i th subset being n_i , is*

$$(4.8) \quad \bar{x} = \frac{1}{N} \sum_{i=1}^k n_i \bar{x}_i, \quad N = \sum_{i=1}^k n_i$$

Corollary. *If $n_i = n$ is the same for all the sets, then $N = kn$ and (4.8) reduces to*

$$(4.9) \quad \bar{x} = \frac{1}{k} \sum_{i=1}^k \bar{x}_i$$

4.7 The Mode. That value of the variable which occurs most frequently in a distribution is called the *mode*. In a sense it is the "fashionable" value (one meaning of "mode" is fashion) and is the kind of average meant in such a phrase as "the average man." If a unimodal distribution can be fitted with a smooth frequency curve, the mode is defined as the abscissa of the highest point on this curve. For the kinds of curves usually fitted, this point can be calculated mathematically, but the calculation requires more advanced mathematics than simple algebra. We shall denote the mode by \bar{x} or sometimes by M_0 .

In a given frequency distribution for a *discrete* variate, the mode can be immediately picked out by inspection. Thus for Table 10 (§2.3) the mode is 2, since the frequency is greater for $x = 2$ than for any other value of x . The difficulty of calculation arises only with a *continuous* variate. We can easily pick out the *modal class*, which for a distribution with equal class intervals is

the class having the greatest frequency,* but the position of the mode within the class is harder to fix.

Sometimes the class mark of the modal class is used, but this is a poor approximation unless the histogram is almost symmetrical. A somewhat better method is illustrated in Fig. 18, applicable when the class intervals in the neighborhood of the mode are all equal.

The method uses three adjacent rectangles of the histogram, with the tallest in the middle. The mode is the abscissa of the point M at which AB and CD intersect. It can be shown that this is also the abscissa of the vertex of a parabola passing through the three points, P , Q , R which are the midpoints of the tops of these three rectangles. If f_{-1} , f_0 , f_1 are the frequencies represented by the three rectangles, c is the class interval, and x_0 is the class mark of the modal class, the formula for calculating the mode is

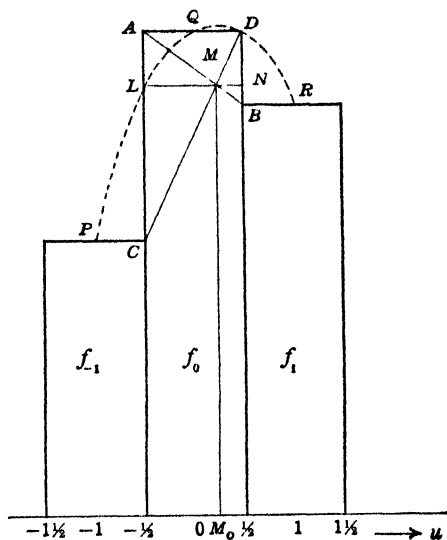


FIG 18 APPROXIMATE MODE

$$(4.10) \quad \hat{x} = x_0 + \frac{c}{2} \frac{f_1 - f_{-1}}{2f_0 - f_1 - f_{-1}}$$

To show this let us change the variate to $u = (x - x_0)/c$. Then the centers of the three class intervals are $u = -1, 0, 1$, respectively. If the abscissa of the mode is \hat{u} , it is evident from the figure that $LM = \hat{u} + \frac{1}{2}$, $MN = \frac{1}{2} - \hat{u}$.

From the geometry of the similar triangles ACM , BDM , we have $\frac{LM}{MN} = \frac{CA}{BD}$

But since the heights of the rectangles are proportional to the frequencies,

$$\frac{CA}{BD} = \frac{f_0 - f_{-1}}{f_0 - f_1}$$

Hence,

$$\frac{\hat{u} + \frac{1}{2}}{\frac{1}{2} - \hat{u}} = \frac{f_0 - f_{-1}}{f_0 - f_1}$$

Clearing of fractions and collecting terms, we have

$$\hat{u}(2f_0 - f_1 - f_{-1}) = (f_1 - f_{-1})/2$$

Finally we return to the x variate by writing $\hat{x} = x_0 + c\hat{u}$, which gives (4.10).

* With unequal intervals, the modal class is that with the greatest frequency per unit of x . In Table 16, it is the class 40-49.

As an illustration we may take Table 12, §2.5, where $x_0 = 134.75$, $c = 10$, $f_{-1} = 196$, $f_0 = 248$, $f_1 = 197$.

By substitution in (4.10) we find $\hat{x} = 134.75 + 5(1/103) = 134.8$ so that the approximate mode is 134.8 lb. It happens in this case that \hat{x} is practically equal to the class mark, since the two frequencies adjacent to the modal class are almost exactly equal.

Sometimes a distribution has two modes, as illustrated in Fig. 19, which represents the distribution of number of petals in flowers of a certain species of chrysanthemum. There are clearly two humps (disregarding small irregularities), one near $x = 23$ and the other near $x = 33$. A distribution like this is often regarded as evidence of heterogeneity in the population. Thus if a frequency curve were drawn for heights of a large sample of adult males, about half American and half Japanese, the curve would similarly be bimodal. It would really be a superposition of two unimodal frequency curves with different modes, the mode for the Americans being higher than that for the

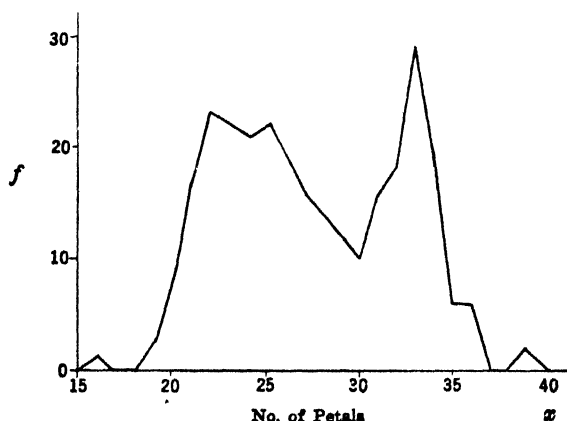


FIG. 19. NUMBER OF PETALS IN *Chrysanthemum leucanthemum*

Japanese. Possibly in the distribution of Fig. 19 we have a mixture of two varieties of the same species, the one tending to have more petals than the other, although in both varieties the number of petals is variable to some extent.

4.8 Relation Between Mean, Median, and Mode. If a distribution is represented by a histogram, an ordinate through the *median* divides the area into two equal parts. An ordinate through the *mean* passes through the centroid of the area; that is to say, if the histogram were cut out of a thin homogeneous metal plate and held in a horizontal plane it would balance about a knife-edge along this ordinate. As already pointed out, an ordinate through the *mode* (if there is only one mode) passes through the highest point of the frequency curve which fits the distribution.

Fig. 20 shows the position of the three averages in a moderately skew distribution. If the distribution were perfectly symmetrical then all three of these measures of location would coincide.

There is an interesting empirical relationship between the three quantities which appears to hold for unimodal curves of moderate asymmetry, namely,

$$\text{mean} - \text{mode} = 3(\text{mean} - \text{median})$$

It is a useful mnemonic to observe that the mean, median, and mode occur in the same order (or reverse order) as in the dictionary; and that the median is nearer to the mean than to the mode, just as the corresponding words are nearer together in the dictionary *

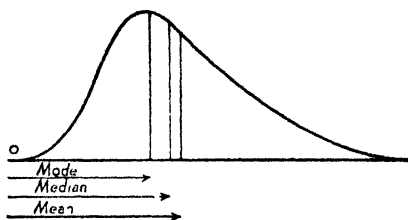


FIG. 20

4.9 Relative Merits of Mean, Median, and Mode. The student primarily interested in the use of these averages in practical statistics might reasonably inquire, "Which of the three averages mentioned should be used in a given problem?" The answer depends upon certain properties peculiar to each average and upon the nature of the data to be averaged.

The qualities desired in an average may be listed as follows. The average should be:

- (a) rigorously defined,
- (b) easily computed,
- (c) capable of a simple interpretation,
- (d) dependent on all the observed values,
- (e) not unduly influenced by one or two extremely large or small values,
- (f) likely to fluctuate relatively little from one random sample to another (of the same size and from the same population),
- (g) capable of mathematical manipulation.

These qualities are satisfied in varying degrees by the averages already described. The most important ones from the standpoint of the ordinary man are (b) and (c), and here the median rates high, but the most important from the point of view of mathematical statistics are (f) and (g). Unfortunately the student at this stage will find it hard to appreciate why this should be so, and will have to take it on trust that the arithmetic mean is distinctly superior in these two, and in most other, respects. A glimpse of what is meant by (g) was given by Theorems 10 and 11. No such simple results as these can be stated for the median or the mode.

* M. G. Kendall, *The Advanced Theory of Statistics*, Lippincott, vol. I, p. 35.

The median has an advantage over the mean in three situations:

(1) When there is an open class at one or both ends of the distribution (as in Table 6, §1.8) the arithmetic mean cannot be calculated but the median can.

(2) When exceptionally large or small values occur at the ends of the distribution, the median may be much more "typical" than the mean. Thus, the median of 41, 43, 46, 48, 49, 52 and 141 is 48 but the mean is 60.

(3) When the observations cannot be measured numerically but can be ranked in order, the middle one (or the two middle ones) can be readily picked out, but the other averages have no meaning.

The mode satisfies condition (c) above, but has little meaning unless the sample is large. It is the appropriate average if we want the "usual" value, the one "most in demand," as in some questions of marketing. It is not, however, easy to compute.

If we are primarily interested in estimating the average of a parent population from the average of a sample (an important practical problem in statistics) the *mean* is usually the proper average to use, as it is more efficient* than any other in the sense suggested by condition (f) above when the population approximates the "normal" type described in Chapter VIII.

4.10 Geometric Mean. The geometric mean of a set of N positive values is the positive N th root of their product. For the N values, x_1, x_2, \dots, x_N ,

$$(4.11) \quad M_G = [x_1 \cdot x_2 \cdots x_N]^{1/N}$$

or

$$(4.12) \quad (M_G)^N = \prod_{i=1}^N x_i$$

Thus the geometric mean of 4 and 9 is 6, which is the positive square root of 36. If any of the x_i are zero, M_G is zero, and if one of them is negative M_G may be imaginary. The geometric mean is never actually used except for positive numbers.

The simplest way of calculating the geometric mean, usually, is to use logarithms. From (4.11)

$$(4.13) \quad \log M_G = [\log x_1 + \log x_2 + \cdots + \log x_N]/N$$

or

$$(4.14) \quad N \log M_G = \sum_{i=1}^N \log x_i$$

Therefore the logarithm of the geometric mean is the arithmetic mean of the logarithms of the values themselves.

*See Reference 2 at the end of the chapter.

Example. Find the geometric mean of 7.90, 13.82, 22.95, 35.34.

Solution.

$$\begin{array}{rcl}
 \log 7.96 & = & 0.90091 \\
 \log 13.82 & = & 1.14051 \\
 \log 22.95 & = & 1.36078 \\
 \log 35.34 & = & 1.54827 \\
 \hline
 & & 4.95047 \\
 \log M_G & = & 1.23762 \\
 M_G & = & 17.28
 \end{array}$$

The geometric mean may not satisfy condition (c) of §4.9, but it is better than the arithmetic mean as regards (e), and, unlike the median, it does depend on *all* the observations. Thus the A.M. of 100, 100, 100, and 1000 is 325 but the G.M. is 177.8. For this reason the geometric mean is often preferred to the arithmetic mean in averaging bacterial counts on agar plates for the purpose of judging milk samples. The bacterial population may easily jump to a very high value in the occasional sample, and the geometric mean is considered the more typical average.

Sometimes when a frequency distribution is graphed it is seen to be quite skew, descending fairly rapidly to the x -axis on the left but stretching out in a long tail to the right. In such a case it often happens that when $\log x$ is used as the variate instead of x , the distribution appears much more nearly symmetrical.* The arithmetic mean of the new variates (which is the logarithm of the geometric mean of the old variates) would be appropriate for this type of distribution.

The geometric mean is the natural average for *ratios*. Suppose we want to compare two values of the ratio of x and y (say the net worth and the debt for two firms), we get an equivalent result by using the geometric mean whether we consider the ratio x/y or the ratio y/x . This is not true for the arithmetic mean. For example in the table

x	y	x/y	y/x
50	20	2.5	0.4
10	8	1.25	0.8

the A.M. of x/y is 1.875 and that of y/x is 0.6, which is not the reciprocal of 1.875. The G.M. of x/y is, however, 1.77 and that of y/x is 0.565, which is the reciprocal of 1.77. Since we feel that the average ought not to depend on the particular way we choose to express the ratio, it seems reasonable to use the geometric mean. Cost-of-living and other index numbers are weighted averages of ratios and here, too, the geometric mean is often used (Chapter V).

4.11 Weighted Geometric Mean. If the observations x_1, x_2, \dots, x_k have weights f_1, f_2, \dots, f_k , the weighted geometric mean is given by

* Several examples are given in an article by J. H. Gaddum, *Nature*, **156**, 1945, pp. 463-466.

$$(4.15) \quad (M_G)^N = x_1^{f_1} x_2^{f_2} \cdots x_k^{f_k}, \quad N = f_1 + f_2 + \cdots + f_k$$

or

$$(4.16) \quad N \log M_G = \sum_{i=1}^k f_i \log x_i, \quad N = \sum_{i=1}^k f_i$$

This arises in finding the average rate of compound interest over a period of years in which a sum of money has accumulated partly at one rate and partly at another. If a sum \$P is invested for n_1 years at $I_1\%$ and then the amount to which it has accumulated is invested for n_2 years at $I_2\%$, the accumulated amount \$A is given by

$$A = P(1 + i_1)^{n_1}(1 + i_2)^{n_2}$$

where $i_1 = I_1/100$, $i_2 = I_2/100$. At the rate $I\%$ for the whole period

$$A = P(1 + i)^{n_1+n_2}, \quad i = I/100$$

and by equating the two values of A we see that $1 + i$ is the weighted geometric mean of $1 + i_1$ and $1 + i_2$.

If the same problem is worked at *simple* interest, we find that i is the weighted *arithmetic* mean of i_1 and i_2 .

4.12 The Law of Growth. The amount of a sum of money at a fixed rate of interest is one example of a quantity which increases according to an exponential law,

$$(4.17) \quad y = ar^x$$

The terms corresponding to $x = 1, 2, 3, \dots$ form a geometric progression with common ratio r . In the compound interest illustration r is equal to $1 + i$. The same law is followed by any quantity which increases at a fixed rate *proportionally to itself*. Since such a rate of increase is characteristic of biological populations with abundant food supplies and no overcrowding (for example, a bacterial culture in its early stages, or the population of a country with a rapidly expanding frontier), this exponential law has been called the *law of growth*.

If we wish to average a set of values taken from a population following approximately such a law, the geometric mean is the proper one to take. Given that the population of a city was 98,000 in 1940 and 129,000 in 1950, and with no further information, having to estimate the population in 1945, the best estimate would be $(98,000 \times 129,000)^{1/2} = 112,000$. The average annual rate of increase is obtained from (4.17) by solving the equation

$$129,000 = 98,000r^{10}$$

which gives $r = 1.047$. The average annual rate of increase is therefore 4.7%. The population for any other year can then be estimated, but great caution

should be exercised about extrapolation outside the decade given, as it can seldom be assumed that the conditions governing the growth of a city remain unaltered for long periods.

4.13 Harmonic Mean. Another average which has long been known and which is required in certain problems is the *harmonic mean* (M_H). For the N positive values x_1, x_2, \dots, x_N , it is defined as the reciprocal of the arithmetic mean of the reciprocals of the values. In symbols,

$$(4.18) \quad M_H = \frac{1}{\frac{1}{N} \left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_N} \right)} = \frac{N}{\sum \left(\frac{1}{x_i} \right)}$$

This measure is used in averaging ratios, such as rates and prices, when certain conditions are agreed upon.

Many ratios can be expressed either as x/y or as y/x . The price of eggs can be put as so many cents per dozen or so many eggs for a dollar. If 100 bushels of wheat are exchanged for 175 dollars, the price of wheat is 175/100 dollars per bushel or 100/175 bushels per dollar. The correct average of such prices will be the arithmetic mean if the unit of the denominator is regarded as fixed and the numerator as variable; but it will be the harmonic mean if the unit of the numerator is regarded as fixed and the denominator as variable.

Suppose we wish to average k ratios $r_i = x_i/y_i$ ($i = 1, 2, \dots, k$). The average is the total amount of x divided by the total amount of y , that is, $\sum x_i / \sum y_i$. First, let us regard the unit of y as fixed and express all the ratios with a common y , equal to v . Then $x_i = r_i y_i = r_i v$. Hence,

$$\sum x_i / \sum y_i = v \sum r_i / kv = (\sum r_i) / k = \bar{r}$$

which is the arithmetic mean of the r_i . If, however, we regard the unit of x as fixed, $y_i = x_i / r_i$, and all the x_i are equal, with a common value u . Then

$$\sum x_i / \sum y_i = ku / [u \sum (1/r_i)] = k / \sum \left(\frac{1}{r_i} \right)$$

which is the harmonic mean of the r_i .

Example 1. A tourist purchases gasoline at three filling stations, where the prices are $33\frac{1}{3}$, 25, and 20 cents per gallon. What is the average price?

Solution. If we regard a gallon as the fixed unit, the average is the arithmetic mean, 26.1 cents/gallon, but if we regard a dollar as the fixed unit, the average is the harmonic mean $3/(3/100 + 1/25 + 1/20) = 25$ cents/gallon. The former would be the correct average price actually paid by the tourist if he bought the same number of gallons at each station. The latter would be the correct average if he spent the same sum (say \$1) at each station. It corresponds to the arithmetic mean of the prices expressed as gallons per dollar, namely, 3, 4, and 5 gallons/dollar.

Example 2. In a certain factory a unit of work is completed by A in 4 minutes, by B in 5 minutes, by C in 6 minutes, by D in 10 minutes, and by E in 12 minutes. What is their average rate of working? At this rate how many units will they complete in a 6-hour day?

Solution. The rates are here expressed as times per unit of work. If we regard the output of work per unit of time as the important consideration, the harmonic mean should be used.

$$M_H = 5 / (\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{10} + \frac{1}{12}) = 300/48 = 6\frac{1}{4}$$

That is, the average rate of working is $6\frac{1}{4}$ minutes per unit. In 360 minutes the five men together will complete $5 \times 360/6\frac{1}{4} = 288$ units of work. This result can be obtained also by considering that the men separately will do 90, 72, 60, 36, and 30 units in 360 minutes, and the sum of these is 288 units. The use of the harmonic mean is therefore justified.

It has been pointed out* that in computing the velocity of a fireball from the path length s and the duration of flight t , by the equation $v = s/t$, the harmonic mean of a number of observations is more natural than the arithmetic mean. This is because s is much better determined than t . The path length is computed by triangulation from the observed positions of beginning and ending, which with experienced observers are fairly well defined. The time is largely a matter of retroactive estimation, and the beginning is difficult to observe accurately because it comes as a surprise. Hence the error is mainly in t , and we are averaging times per unit of distance rather than distances per unit of time.

4.14 Relation of Arithmetic, Geometric, and Harmonic Means. The geometric mean occupies a middle-of-the-road position between the arithmetic and harmonic means. It is not as sensitive as the A.M. to a few very high values or as sensitive as the H.M. to a few very low ones. Also, for any set of positive numbers,

$$A.M. \geq G.M. \geq H.M.$$

This is easily proved for *two* numbers, say a and b . If $a = b$ the three means are all equal, so that we may take $a > b$. Denoting the means temporarily by A, G, H , we have

$$A = (a + b)/2, \quad G = \sqrt{ab}, \quad H = 2ab/(a + b)$$

Now $(\sqrt{a} - \sqrt{b})^2 > 0$, so that, squaring out,

$$a + b - 2\sqrt{ab} > 0$$

whence

$$(4.19) \quad (a + b)/2 > \sqrt{ab}$$

or

$$A > G$$

* W. J. Fisher, *Harvard College Observatory Circular No. 375*, 1932.

Again, on multiplying (4.19) through by $2\sqrt{ab}/(a+b)$, we obtain

$$\sqrt{ab} > 2ab/(a+b)$$

or

$$G > H$$

Hence for any two positive numbers a and b ,

$$A \geq G \geq H$$

(As a mnemonic, this is the order of the letters in the alphabet.)

It is a little more difficult to prove the relation for N numbers, x_1, x_2, \dots, x_N . By definition,

$$A = (x_1 + x_2 + \dots + x_N)/N$$

$$G^N = x_1 \cdot x_2 \cdot \dots \cdot x_N$$

$$N/H = 1/x_1 + 1/x_2 + \dots + 1/x_N$$

If all the x , are equal, $A = G = H$. If not, take the least (x) and the greatest (x_k) and replace each by $(x + x_k)/2$. (There may, of course, be several numbers all equal to x , or x_k ; we take any one of them.) This process leaves A unaltered but changes G to G_1 , replacing x, x_k in G by $(x + x_k)^2/4$, which by (4.19) is greater than x, x_k . Hence $G_1 > G$. Also the numbers are now nearer to each other than they were before. We continue this process as long as any x , remain unequal, obtaining a sequence of values $G < G_1 < G_2 < G_3 < \dots$ and at each step the range of the N numbers is equal to or less than before. The limit of the process is a set of N numbers all equal to each other and to A . Therefore $G < A$.

The process may terminate in a finite number of steps or the limit may never actually be reached. In the latter case, however, we can continue as long as we like. Suppose $G_1 - G = d$, which is positive. Let us continue until the difference between the greatest and least of the N numbers is less than d . The difference between the geometric mean (say G_n) and the arithmetic mean (A) of these numbers will certainly be less than d , and $G_n - G > d$. Hence $G < A$.

A similar argument may be used to prove that $G > H$. We replace $1/x$, and $1/x_k$ by $(1/x + 1/x_k)/2$. The details are left to the student. The proof is not important in itself, but it provides drill in careful reasoning.

4.15 Root Mean Square. Another average that is sometimes used is the *root mean square* (R.M.S.) defined as the positive square root of the mean of the squares of the values x , that is,

$$(4.20) \quad \text{R.M.S.} = (\overline{x^2})^{1/2} = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right)^{1/2}$$

The R.M.S. can be used when the x , are sometimes positive and sometimes negative. It is, for example, the average used for an alternating electric

current. If values were measured for an alternating current at many instants throughout a time covering many oscillations, the arithmetic mean would be near zero. The effective current from the point of view of the consumer is the root mean square current.

In theoretical statistics the chief application of the root mean square is the measurement of deviations of the values x_i of a variate from the mean \bar{x} . The R.M.S. of these deviations is called the *standard deviation* and is by far the most important measure of dispersion (Chapter VI).

Exercises

1. Write in expanded form:

$$(a) \sum_{i=1}^k x_i^2 f_i$$

$$(b) \sum_{i=1}^k (x_i - \bar{x}) f_i$$

$$(c) \sum_1^{n_1} f_i$$

$$(d) \sum_{n_1+1}^{n_1+n_2} x_i f_i$$

$$(e) \prod_{i=1}^N (x_i + c)$$

2. Write in abbreviated notation:

$$(a) x_1 f_1 + x_2 f_2 + \dots + x_k f_k$$

$$(b) \frac{1}{N} [(x_1 - \bar{x})^2 f_1 + (x_2 - \bar{x})^2 f_2 + \dots + (x_k - \bar{x})^2 f_k]$$

$$(c) (a+1)(a+2)(a+3) \dots (a+N)$$

$$(d) a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$$

3. Prove:

$$(a) \sum_{i=1}^N (x_i - c) = \sum_{i=1}^N x_i - Nc$$

$$(b) \sum_{i=1}^k (x_i + 1)^2 f_i = \sum_{i=1}^k x_i^2 f_i + 2 \sum_{i=1}^k x_i f_i + N$$

$$(c) \sum_{x=0}^n x(x-1)p = \sum_{x=2}^n x(x-1)p$$

4. For the example of §4.4, compute $\sum_i (x_i - \bar{x}) f_i$, using the following form:

$$\begin{array}{cccc} x_i & f_i & (x_i - \bar{x}) & (x_i - \bar{x}) f_i \\ 5 & & & \\ 6 & & & \\ 7 & & & \\ 8 & & & \end{array}$$

5. Show by writing out the sums that $\sum_{i=1}^N x_i y_i$ is not the same as $\left(\sum_{i=1}^N x_i \right) \left(\sum_{i=1}^N y_i \right)$.

6. Use the identity $x^2 - (x-1)^2 = 2x - 1$ to give a proof of Theorem 4 on the lines of the proof given for Theorem 5.

7. Show that the arithmetic mean of the first N integers is $(N+1)/2$.

8. Rewrite Table 15, §4.5, using $x_0 = 74.5$, and verify that \bar{x} is unaltered.

9. Find the arithmetic mean of 18, 42, 23, 16, 103, 61, 49, 95, 113, 10, using (4.1).

Find the deviations of each of these numbers from 50 and verify that the mean of these deviations, plus 50, gives the mean of the numbers themselves.

10. Compute by the fully coded method of §4.5 the arithmetic mean of the weights of 1000 students (Table 12, §2.5). *Ans.* 138.65 lb.

11. From Table 5, §1.8, find the mean monthly rainfall at Iowa City for the 36 years given.

12. Find the mean of the distribution of the discrete variate in Table 11 (§2.3). *Ans.* $\bar{x} = 6.139$.

13. The mean grade of one class of 20 students is 66% and that of another class of 15 students is 70%. Find the mean grade of the two classes taken together.

14. Calculate the mean income of Canadian taxpayers from the data of Table 6, §1.8. (Note that this cannot be done without further information because of the open classes at the ends. Assume for the purpose of this question that the first class starts at \$1000 and that the last class ends at \$250,000.)

15. Find the mean of the following distribution:

x	f
47.5	7
48.1	17
45.9	46
44.0	44
40.7	54
41.6	43
38.0	35
33.2	14

16. In chemistry a student was graded 85 in class work, 80 in laboratory, and 65 in final examination. If these were weighted 1, 2, and 3 respectively, what was the student's average grade?

17. The population of a city increased in 5 years from 225,000 to 245,000. What was the average annual rate of increase, assuming that the "law of growth" applied?

18. The number of bacteria in a certain culture was found to be 4×10^6 at noon of one day. At noon the next day the number was 9×10^6 . If the number increased at a constant rate per hour, how many bacteria were there at the intervening midnight?

19. For five successive years the rates of interest on money were 4.25, 5.30, 4.65, 3.86, and 4.38%. What was the average rate of interest? (Use the G.M. of $1 + r$.)

20. Show that if a sum of money \$ P accumulates at simple interest for n_1 years at $I_1\%$ and for n_2 years at $I_2\%$, the average rate of interest is $(n_1 I_1 + n_2 I_2)/(n_1 + n_2)$. (Cf. §4.11.)

21. The following table gives the population of the United States at each 10-year census from 1860 to 1940. (*Historical Statistics of the United States*, 1949.)

Year	Population (Millions)	Ratio to Preceding Figure
1860	31.4	
70	39.8	1.27
80	50.2	1.26
90	62.9	1.25
1900	76.0	1.21
10	92.0	1.21
20	105.7	1.15
30	122.8	1.16
40	131.7	1.07

What is the average rate of increase per decade? Estimate the population of 1950 from the 1940 figure. *Ans.* 1.20, 156.7 millions. (The true value in 1950 was 150.7 millions.)

22. Given two sets, each of n positive values, $x_{11}, x_{12}, \dots, x_{1n}, x_{21}, x_{22}, \dots, x_{2n}$, prove that the geometric mean of the ratios of corresponding values in the two sets is equal to the ratio of the geometric means of the two sets.

23. If x_1 and x_2 are two positive values of a variate, prove that their geometric mean is equal to the geometric mean of their arithmetic and harmonic means.

24. (*Amer. Math. Monthly*, 42, 1935, p. 394) Show that if $2a$ is the harmonic mean of two rational numbers b and c , then the sum of the squares of the three numbers a, b, c is the square of a rational number.

25. Find to four significant figures the arithmetic, geometric, and harmonic means of the first 15 positive integers. Verify the relationship $A > G > H$.

26. (*Burgess*). Twenty boats make 6 transatlantic trips each per year, and 10 boats make 4 trips each per year. What is the average number of days for a "turn around" (that is, time between consecutive departures from the same port)? Take the year as 360 days for convenience.

Hint. The rates may be expressed either as trips per year (6 or 4) or as days per trip (60 or 90), and the time unit may be regarded as fixed. The first method requires the arithmetic mean and the second the harmonic mean. Both give the same result, $5\frac{1}{2}$ trips per year or 67.5 days per trip.

27. A plane travels one-half of a given distance, D miles, at a speed of x_1 miles per hour, and the remaining half distance at a speed of x_2 miles per hour. Show that the average speed for the entire distance is the harmonic mean of x_1 and x_2 . (If x_1 and x_2 are rates for going and returning, half the average speed is the "radius of action per hour," that is, the distance that the plane could travel and return in one hour.)

28. Three boys work correctly 7, 10, and 15 arithmetic problems during a half-hour test. Assuming that the problems are of about the same difficulty, what is the average rate at which the boys work?

Hint. Is the significant measure of speed of working (a) the time a boy takes to work one problem or (b) the number of problems he can work in a given time? If you prefer (a), use the harmonic mean, if (b) the arithmetic mean. Authorities differ as to which is the more reasonable interpretation.

29. Write out the details of the proof in §4.14 that $G > H$ for N numbers.

30. With a large amount of data, the calculation of the geometric mean may be simplified by coding, as is done for the arithmetic mean. Show that if we choose class intervals with a constant ratio of the upper to the lower boundary, and form a frequency table, the coded method of §4.5 applies, with

$$\log M_G = x_0 + \frac{c}{N} \sum f_i u_i$$

when x_0 is midway between the *logarithms* of the upper and lower boundaries of the selected class interval, and c is the logarithm of the ratio between these boundaries. (See Reference 3.)

31. Apply the method of Exercise 30 to the data of Table 2, §1.8. (Here the highest grade is 98 and the lowest 34, with a ratio of 2.88. If we want 7 classes the ratio of class boundaries should be $(2.88)^{1/7} = 1.16$. If we take the lowest class boundary arbitrarily as 32.0, the succeeding ones are 37.1, 43.1, 49.9, 57.9, 67.2, 77.9, 90.4, approximately.) Form the frequency distribution and complete the calculation of the geometric mean.

32. Prove that for any set of numbers x_1, x_2, \dots, x_N the arithmetic mean is not greater than the root mean square.

Hint. Prove this first for *two* numbers.

References

1. Yule and Kendall, (see our §0.4), Ch. VII.
2. W. A. Shewhart, *Economic Control of Quality of Manufactured Products*. (D. Van Nostrand Co., Inc., 1931), p. 280.
3. Z. Szatrowski, "Calculating the Geometric Mean from a Large Amount of Data," *J. Amer. Stat. Assoc.*, **41**, 1946, pp. 218–220.
4. W. F. Fergar, "The Nature and Use of the Harmonic Mean," *J. Amer. Stat. Assoc.*, **26**, 1931, pp. 36–40.

CHAPTER V

INDEX NUMBERS

5.1 Index Numbers as Weighted Averages. The index number is a widely used statistical device for comparing one group of related variables with another group. We may wish to compare the prices of many different articles of food at one date with the prices of the same articles at a different date, and to express the comparison by a single food-price index number. Or we may compare the production of a group of farm crops in one country with that of the same crops in another country, obtaining a single crop-production index for the one country as compared with the other. Or, again, we may compare the scores of two individuals on a battery of tests, expressing their relative scores by a single number. All of these are examples of *index numbers*, and it is evident that they are really averages of ratios. The many different price ratios for beef, pork, bread, sugar, etc., for example, are averaged to give the single price-index number for food. Moreover, the average must be a *weighted* average, because not all the items included in the group are equally important. In the construction of a food-index number it would clearly be unreasonable to give pepper and salt the same weight as bread and beef.

In order to compute an index number it is necessary to collect a mass of data on prices, production, scores, or whatever is being compared. But it is also necessary to decide on the type of average to be used and the relative weighting of the different items in the group, and here there is room for considerable differences of opinion. Arithmetic, geometric, and harmonic means all have their advocates, and several different methods of weighting have been suggested. We shall describe some of these, as illustrations of the various averages described in Chapter IV, but we shall not enter into the important practical details of how the data are actually obtained, how accurate they are, or what governs the choice of items that make up the group.

A related set of index numbers, such as a series of price-index numbers computed over a period of months or years, is called an *index*. The most familiar example is the cost-of-living index, computed monthly by the United States, Canadian, and other governments.

5.2 Price Index Numbers. Let us suppose that we wish to compute a price index for nonferrous metals in the United States over a number of years, and that we have the data from official sources on the prices of copper, zinc, lead, aluminum, etc., over the period studied. We must choose a base year such as 1939 with which the other years are to be compared, and we will call this the year 0. Let the price of the copper be $p_0^{(1)}$ in the base year and $p_n^{(1)}$

in the year n , which is one of the years for which we want the index number. The quantity $p_n^{(i)}/p_0^{(i)}$ is called a *price relative*. It expresses the price in year n as a fraction of the price in year 0, and it is usually multiplied by 100 to make it a percentage. Similarly the price relative for zinc would be $p_n^{(2)}/p_0^{(2)}$ and so for all the N metals making up the selected group of non-ferrous metals. A simple arithmetic mean of these price relatives would give a price index number, but such a procedure would give equal importance to all the metals in the list. Instead it is customary to weight the price relatives according to the *value* of metal produced, either in the base year or in the current year, in both cases using base year prices.

Let $q_0^{(i)}$ be the quantity of copper produced in year 0, $q_n^{(i)}$ the quantity produced in year n , and so on for the other metals. Then $v_0^{(i)} = p_0^{(i)}q_0^{(i)}$ will be the value of the copper produced in year 0, and $v_n^{(i)} = p_n^{(i)}q_n^{(i)}$ that of the production in year n at the price of year 0. Thus if $p_0^{(i)}$ is in dollars per ton and $q_0^{(i)}$ is in tons, $p_0^{(i)}q_0^{(i)}$ will be in dollars. Laspeyres' index number uses $v_0^{(i)}$ as the weight for the i th item (that is, the value in the base year) and Paasche's index number uses $v_n^{(i)}$ (the value in the current year at base year prices).

$$(5.1) \quad P_L = \frac{\sum_i v_0^{(i)} p_n^{(i)} / p_0^{(i)}}{\sum_i v_0^{(i)}} = \frac{\sum p_n^{(i)} q_0^{(i)}}{\sum p_0^{(i)} q_0^{(i)}}$$

$$(5.2) \quad P_P = \frac{\sum_i v_n^{(i)} p_n^{(i)} / p_0^{(i)}}{\sum_i v_n^{(i)}} = \frac{\sum p_n^{(i)} q_n^{(i)}}{\sum p_0^{(i)} q_n^{(i)}}$$

The symbol P is used since both these index numbers refer to *prices*. The subscripts L and P refer to Laspeyres and Paasche.

The advantage of (5.1) is that once the weights have been determined they stay fixed, and it is only the prices that change from year to year. In (5.2) the weights have to be computed afresh for each value of n . It may well be that if conditions of production are changing, the current quantities give a more realistic picture than the base-year quantities, but, on the other hand, the base year is usually selected as a typical year when conditions could be regarded as approximately normal, and the use of base-year quantities gives more stability. Sometimes an average over several years is used as a base. The Canadian cost-of-living index was based on an average of the years 1935 to 1939.* The number P_L measures the change in cost, due to change in prices, of a fixed set of commodities in fixed amounts. It tends to be too *high* as an index of prices because, in a market which is to some extent free, consumers will tend to reduce their purchases of more expensive items and increase those of less expensive ones. The number P_P measures the change in

* In 1952 a new consumer price index, based on 1949, was instituted.

cost of the quantities actually purchased from what the same quantities would have cost had they been purchased in the base year. The same argument about a changing market shows that the Paasche index tends to be too *low*, because the ratio is always in the direction current year to base year, and the current consumption might overestimate the total of expenditure when applied to base year prices. This does not mean that P_L is always greater than P_P as actually computed — frequently it is not so. The numbers relate to different bases of reference.

To compromise between the two points of view we may use as weights the *arithmetic mean* of $v_0^{(i)}$ and $v_n^{(i)}$, as in the Marshall-Edgeworth formula, or the *geometric mean*, as in the Walsh formula (the names are those of writers who have prominently advocated these formulas). Dropping the superscript i , which is however to be understood, we may write these index numbers as

$$(5.3) \quad P_{ME} = \frac{\sum (v_0 + v_n) p_n / p_0}{\sum (v_0 + v_n)} \\ = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)}$$

$$(5.4) \quad P_W = \frac{\sum \sqrt{v_0 v_n} p_n / p_0}{\sum \sqrt{v_0 v_n}} \\ = \frac{\sum \sqrt{q_0 q_n} p_n}{\sum \sqrt{q_0 q_n} p_0}$$

Another compromise is to compute both P_L and P_P and take the mean, either arithmetic or geometric. The former was recommended by Bowley and the latter by Irving Fisher, who called it the “ideal” index number.

$$(5.5) \quad P_B = (P_L + P_P) / 2$$

$$(5.6) \quad P_F = \sqrt{(P_L P_P)}$$

The formula that is perhaps the most frequently-used practical compromise is the *fixed-weight aggregative*. The weights are the values at base-year prices of quantities $q_a^{(i)}$ which represent neither the actual base-year consumption (or production) nor the actual current-year consumption, but measure in some way the estimated relative importance of the items. Mitchell has advocated this formula with weights which are the average quantities of the commodities bought and sold over a period of years.

$$(5.7) \quad P_M = \frac{\sum v_a p_n / p_0}{\sum (v_a)}, \quad v_a^{(i)} = p_0^{(i)} q_a^{(i)}$$

or in aggregative form

$$P_M = \sum (p_n q_a) / \sum (p_0 q_a)$$

As we shall see later, the Fisher index number has some technical advantages, although it is harder to compute, and to interpret when computed, than most of the others. The Marshall-Edgeworth index number is regarded as a good, practical, all-round formula. The Mitchell index-number suffers from the fact that the fixed weights tend to become out of date.

With any index number it is useful to express the weights as fractions of the total so that they add up to unity. For example, if $v_0' = v_0/\sum v_0$,

$$(5.8) \quad P_L = \sum v_0'(p_n/p_0),$$

so that the index number is a sum of numbers each corresponding to one item in the group of commodities which forms the basis of the index. The contribution of each item toward the whole index number is therefore evident at a glance.

5.3 An Example. The computation of a useful index number is likely to be laborious because of the large number of commodities that need to be included and the variety of prices (for different qualities, styles, etc.) within a single commodity. About 900 commodities and about 1800 price quotations are used in making up the United States Bureau of Labor Statistics Index of Wholesale Commodity Prices, and this index number is computed every week. Purely for the purposes of illustration, we take approximate figures for Canadian prices and per capita consumption for a few selected food items in 1939 and 1949. The data are presented in Table 17, where 1939 is reckoned as the base year and 1949 the current year for which an index number is to be calculated. The calculation of the Laspeyres and Paasche

TABLE 17. APPROXIMATE CANADIAN PRICES (CENTS PER POUND) AND PER CAPITA CONSUMPTION (POUNDS) OF SELECTED FOOD ITEMS, 1939 AND 1949

<i>Food</i>	p_0 (1939)	p_n (1949)	q_0 (1939)	q_n (1949)
Flour	3 3	6 9	185	152
Potatoes	2 25	3 31	192	191
Veal	17 4	55.7	10 5	10.9
Sugar	6 4	9 7	94 7	98.1
Coffee	34 5	65 0	3 7	6.8
Cheese (Cheddar)	26 8	61.4	3.4	3.3
Breakfast food	18 6	30 1	7 4	5.9

(Adapted from data in *Canada Year Book*, 1950, and *Labour Gazette*, 1952)

Note. The published figures of prices and consumption are difficult to compare. They generally refer to different classifications of the items, or to different periods of time. The figures above are crude approximations.

TABLE 17a. CALCULATION OF INDEX FOR ITEMS OF TABLE 17

<i>Food</i>	p_n/p_0	v_0	$v_0' = v_0/\sum v_0$	$v_0' p_n/p_0$	v_n	$v_n' = v_n/\sum v_n$	$v_n' p_n/p_0$
Flour	2.09	611	0.279	0.583	502	0.230	0.481
Potatoes	1.47	432	0.197	0.290	430	0.197	0.290
Veal	3.20	183	0.084	0.269	190	0.087	0.278
Sugar	1.51	606	0.277	0.418	628	0.288	0.435
Coffee	1.89	128	0.058	0.110	235	0.108	0.204
Cheese	2.29	91	0.042	0.096	88	0.040	0.092
(Cheddar)							
Breakfast food	1.62	138	0.063	0.102	110	0.050	0.081
Total		2189	1.000	1.868	2183	1.000	1.861

$$P_L = 1.87 \text{ or } 187\%$$

$$P_P = 1.86 \text{ or } 186\%$$

index numbers is shown in Table 17a, and it is evident that these are so close together that it is hardly worth while to work out the index numbers which will lie between them.

5.4 Quantity Index Numbers. Instead of comparing prices we may be interested in comparing the production (or consumption) from year to year of the commodities making up the group. That is, we want to average *quantity relatives*, $q_n^{(i)}/q_0^{(i)}$, and these may be weighted, like the price relatives, in various ways. We can, in fact, form seven quantity index numbers analogous to the seven price index numbers already mentioned, for example,

$$(5.9) \quad Q_L = \frac{\sum w_0(q_n/q_0)}{\sum w_0} = \frac{\sum p_0 q_n}{\sum p_0 q_0}$$

$$(5.10) \quad Q_P = \frac{\sum w_n(q_n/q_0)}{\sum w_n} = \frac{\sum p_n q_n}{\sum p_n q_0}$$

$$(5.11) \quad Q_F = \sqrt{Q_L Q_P} = \left[\frac{\sum p_0 q_n \cdot \sum p_n q_n}{\sum p_0 q_0 \cdot \sum p_n q_0} \right]^{1/2}$$

Here the weights $w_0^{(i)}$ are the values of base year consumption at base year prices, and thus are the same as the $v_0^{(i)}$, but the weights $w_n^{(i)}$ are the values of base year consumption at current year prices (that is, $p_n^{(i)} q_0^{(i)}$) and therefore are not the same as the $v_n^{(i)}$.

5.5 Fisher's Tests for Index Numbers. Irving Fisher suggested that a good index number should satisfy two tests which he called the "time-reversal test" and the "factor-reversal test." The first means that the number for year n relative to year 0 should be the reciprocal of the number for year 0 relative to year n . This is obviously true for *one* commodity, since p_n/p_0 is

the reciprocal of p_0/p_n , but it is not true for some index numbers, as is easily seen by interchanging 0 and n in the expressions (5.1) and (5.2). Thus, $\sum p_0 q_n / \sum p_n q_n$ is not the reciprocal of $\sum p_n q_0 / \sum p_0 q_0$. The test is satisfied, however, by P_{MF} , P_W , and P_F .

The second test means that a formula which is right for prices should also be right for quantities. That is, the price index multiplied by the corresponding quantity index should give the true ratio of value in year n to value in year 0. This is true for a single commodity, since p_n/p_0 times q_n/q_0 gives $p_n q_n / p_0 q_0$, but it is not true for most index numbers. Thus,

$$P_L Q_L = \frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum q_n p_0}{\sum q_0 p_0}$$

and this is not equal to the ratio of values $\sum p_n q_n / \sum p_0 q_0$. The test is, of course, satisfied by Fisher's index number.

A good test for the consistency of a calculated price index number is the difference between P_L and P_P . If this difference is small, say less than 2 points, either of them, or a mean of the two, may be accepted as a reasonable measure of price change, but if the difference is large not much confidence can be placed in any index number for the data used.

5.6 Geometric and Harmonic Means of Relatives. It has been stated in §4.10 that the geometric mean is a suitable average for ratios, and since price and quantity relatives are ratios the geometric mean (suitably weighted) should give a good index number. It has, in fact, been extensively used in Britain. If we use the weights v_0 (base year values), the geometric mean index number is given by

$$(5.12) \quad (P_{GM})^{2v_0} = (p_n^{(1)} / p_0^{(1)})^{v_0^{(1)}} \cdot (p_n^{(2)} / p_0^{(2)})^{v_0^{(2)}} \cdots (p_n^{(N)} / p_0^{(N)})^{v_0^{(N)}} \\ = \Pi (p_n / p_0)^{v_0}$$

or, writing $v_0 / \sum v_0 = v_0'$,

$$(5.13) \quad \log P_{GM} = \sum v_0' \log (p_n / p_0)$$

This is a reliable index, and one which is comparatively little affected by sampling error (the error caused by the selection of the particular items to make up the index.) It does not satisfy in general the two Fisher tests described in §5.5.

Ferguson (see Ref. 4 of Chap. IV) has argued for the harmonic mean,

$$(5.14) \quad P_{HM} = \frac{\sum v_0}{\sum (v_0 p_0 / p_n)} = \frac{\sum (p_0 q_0)}{\sum (p_0^2 q_0 / p_n)}$$

on the ground that the harmonic mean is lower than either of the others and so takes account of the economies that may be practiced by suitably varying the relative quantities of the different commodities purchased. These economies will lower the effective price level.

5.7 Series of Index Numbers. A long series of index numbers may be calculated either by using the same base period all the time (the *fixed-base* method) or by linking together sets of index numbers for successive periods (the *chain* method). If we denote the base period by 0 and the successive following periods by 1, 2, 3, ..., $n \dots$ the series of fixed-base index numbers may be written

$$P_{01}, P_{02}, P_{03}, \dots, P_{0n} \dots$$

each calculated by direct application of one of the formulas given above. The successive chain index numbers may be written

$$P_{01}, P_{01} \cdot P_{12}, P_{01} \cdot P_{12} \cdot P_{23} \dots$$

For the fixed weight aggregative formula the two series are equivalent. Thus

$$P_{02} = \sum p_2 q_a / \sum p_0 q_a$$

$$P_{01} \cdot P_{12} = \frac{\sum p_1 q_a}{\sum p_0 q_a} \cdot \frac{\sum p_2 q_a}{\sum p_1 q_a} = P_{02}$$

But for the better formulas this is not true. With Laspeyres' index number,

$$P_{02} = \sum p_2 q_0 / \sum p_0 q_0$$

$$P_{01} \cdot P_{12} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \cdot \frac{\sum p_2 q_1}{\sum p_1 q_1}$$

and $P_{01} \cdot P_{12}$ is not the same as P_{02} . The fixed-base system is the easier to calculate, but the chain system is more reliable, for it uses weights which are calculated afresh for each link in the chain.

Sometimes it is convenient to change one commodity for another in the index. In the textile world new fibers (e.g., nylon) are continually coming into use, and old ones (e.g., real silk) practically dropping out of use. It is possible to change the one for the other, provided we have data for an overlapping period, without disrupting the index. Thus, if $q_0^{(1)}$, $q_0^{(2)}$ are base year quantities for commodities X_1 and X_2 , the value of $q_0^{(2)}$ being hypothetical because X_2 did not exist in the base year, and if $p_n^{(1)}$, $p_n^{(2)}$ are current year prices for the two commodities, we may put

$$q_0^{(2)} = q_0^{(1)} p_n^{(1)} / p_n^{(2)}$$

This makes $p_n^{(2)} q_0^{(2)} = p_n^{(1)} q_0^{(1)}$, so that for the overlapping period it makes no difference whether we use X_1 or X_2 in the index.

5.8 Adjusted Death Rates. Death rates are expressed as number of deaths per 1000 of population, and naturally these rates differ in different age groups, being largest among the very young and the very old. In comparing the death rates of two communities it is not accurate to use simply the crude death rates regardless of age, because the two communities may

have a different age composition. (One may have a larger proportion of young children than the other, for instance.) A method of allowing for this is to form for each community a weighted average of the specific death rates in each age group, using the same weights for both communities, namely, the numbers in each age group in a certain fixed standard population. The weighted average of specific death rates is thus a kind of *index-number* for the community and may fairly be used in a comparison. It is called a *standardized* or *adjusted death rate*. In *Statistical Abstracts of the United States*, for the last few years, specific death rates are given for various age groups, as well as an adjusted death rate. For the adjustment a standard population based on the actual United States population of 1940 is used.

A simple example of this comparison of death rates is given by Pearl (Reference 3). The crude death rates in 1910 of two cities of about the same population (Providence, R. I., and Seattle, Wash.) were 17.76 and 10.26, with a ratio of 1.73. The specific rates for different age groups of the population are given in Table 18, together with a "standard million" based on the

TABLE 18. DEATH RATES FOR PROVIDENCE, R. I., AND SEATTLE, WASH., 1910

<i>Providence (Population 224,050)</i>				
Age Group	Percentage of Total Population	Specific Death Rate (p)	Standard Million (q)	$pq/10^6$
0-5	9.74	53.86	115,806	6.24
5-10	8.35	3.96	106,321	0.42
10-20	17.10	3.76	197,931	0.74
20-40	37.30	7.13	333,379	2.38
40-60	20.75	18.37	178,845	3.28
60-80	6.30	67.61	62,391	4.22
over 80	0.47	172.02	5,327	0.92
	100.01		1,000,000	18.20
<i>Seattle (Population 234,719)</i>				
0-5	7.26	26.58	115,806	3.08
5-10	6.44	3.31	106,321	0.35
10-20	13.92	3.28	197,931	0.65
20-40	46.58	5.70	333,379	1.90
40-60	21.22	12.55	178,845	2.24
60-80	4.32	44.08	62,391	2.75
over 80	0.25	174.58	5,327	0.93
	99.99		1,000,000	11.90

(0-5 means up to but not including 5 years)

actual composition of the United States population in 1910, which gives the numbers, out of a total of one million, in each of the age groups cited. The adjusted death rates are then calculated as weighted averages of specific death rates and the results are 18.20 for Providence and 11.90 for Seattle, with a ratio of 1.53. The apparently high ratio of death rates for Providence and Seattle was therefore due, at least in part, to differences in age composition of the two populations, and in fact a breakdown of the population figures shows that Seattle had in 1910 a considerably higher proportion than Providence of young adults (a group with low specific death rate) and a considerably lower proportion of children under 5 and adults over 60 (groups with high specific death rates).

Exercises

1. Compute from Table 19 four index numbers (Laspeyres, Paasche, Marshall-Edgeworth, and Fisher) of farm crop prices in the United States for 1935, with 1919 as base year.
Ans. $P_L = 38.1$, $P_P = 37.5$, $P_F = 37.8$, $P_{ME} = 37.8$ (1919 price = 100).

TABLE 19. AVERAGE PRICES AND TOTAL PRODUCTION FOR TWELVE LEADING FARM CROPS IN UNITED STATES, 1919 AND 1935

Crop	Unit	Price (dollars)		Production (millions)	
		1919	1935	1919	1935
Corn	bus.	1.343	0.547	2679	2203
Wheat	bus.	2.131	0.894	952.1	603.2
Cotton	lb.	0.356	0.115	5705	5365
Hay	ton	20.15	7.23	76.59	75.62
Oats	bus.	0.702	0.257	1107	1195
Tobacco	lb.	0.390	0.185	1444	1284
Potatoes	bus.	1.580	0.634	297.3	356.4
Sugar	lb.	0.102	0.031	4371	*6278
Barley	bus.	1.215	0.377	131.7	292.2
Rye	bus.	1.331	0.402	78.7	57.9
Rice	bus.	2.666	0.624	42.69	38.45
Flaxseed	bus.	4.383	1.548	6.77	14.93

Source: *Yearbook of Agriculture*, and "Crops and Markets" (U. S. Dept. of Agriculture). Quoted by Mills, *Statistical Methods*, 1938 (Holt), pp. 179-206.

* 1934 production figure used.

2. Draw up a table of logarithms of price relatives for the data of Table 19, and compute a geometric mean index of farm prices, using weights based on 1919 production.

Ans. $P_{GM} = 37.8$.

3. Table 20 gives the average prices of selected types of meat at three periods, April 1940, January 1947, and January 1948, in retail stores in Edmonton, Alberta, and also average quantities of each purchased per family by housewives in the month of April 1940.

Construct a retail meat price index for 1947 and 1948 (Laspeyres' formula), with 1940 as base year. *Ans.* $P_L = 166.5$ (1947), 196.9 (1948).

TABLE 20. PRICES OF SELECTED CUTS OF MEAT, EDMONTON, 1940, 1947, 1948
AND CONSUMPTION, 1940

<i>Meat</i>	$p_0(1940)$ cents/lb	q_0 lb	$p_1(1947)$	$p_2(1948)$
Beef, prime rib roast	18.48	10	30.46	35.58
Beef, rolled rib roast	23.00	5	40.27	44.50
Beef, shoulder roast	13.71	4	24.18	30.07
Beef, flank stew	9.72	4	19.19	24.04
Beef, sirloin steak	22.30	2	43.76	47.85
Veal, shoulder roast	15.46	4	23.15	27.56
Pork, shoulder roast	17.38	2	27.33	36.61
Pork, loin	24.83	0.25	40.00	45.62
Bacon, sliced side	29.90	1.75	50.74	65.71
Ham, boiled sliced	54.31	0.75	69.13	78.69

4. Table 21 gives approximate figures for the specific death rates by race for Mississippi, Florida, and the United States, 1948. The crude death rates for these two states were 9.9 and 9.5, per 1000. The United States population as a whole may be taken as 89.8% white, and Mississippi and Florida as about 50.7% and 72.8% white. Calculate adjusted death rates for Mississippi and Florida, corrected for the different race composition of the two states. (Use the actual United States population as standard.) *Ans.* 9.3 and 9.5.

TABLE 21. SPECIFIC DEATH RATES (PER 1000) BY RACE, 1948

	<i>White</i>	<i>Non-White</i>
United States	9.7	11.7
Mississippi	9.1	10.7
Florida	9.4	10.3

Source: *Statistical Abstracts of U. S.*, 1951.

5. Adjust the death rates for Seattle and Providence (Table 18) by the following "standard million," which is based on the population of England and Wales in 1901 and which has been widely used.

ENGLAND AND WALES STANDARD MILLION

<i>Age Group</i>	<i>Population</i>	<i>Age Group</i>	<i>Population</i>
0-4.....	114,262	40-44.....	56,893
5-9.....	107,209	45-49.....	48,365
10-14.....	102,735	50-54.....	40,857
15-19.....	99,796	55-59.....	32,359
20-24.....	95,946	60-64.....	27,382
25-29.....	86,833	65-69.....	19,358
30-34.....	74,746	70-74.....	13,722
35-39.....	65,956	75-79.....	8,131
		80-.....	5,450
			<u>1,000,000</u>

6. (*Woods and Russell*) Specific death rates for clergymen and for railwaymen (1900-1902) are given in the following table:

Age group	Clergymen	Railwaymen
45-54	9.82	13.34
55-64	23.38	29.76
65-	82.57	93.17

The crude death rates were 35.34 for clergymen and 26.52 for railwaymen, with a ratio of 1.33. Show that when the rates are adjusted to the England and Wales standard million (Exercise 5), they become 31.31 and 37.39, with a ratio of 0.837. (The great difference between adjusted and unadjusted rates is due to the fact that the proportion of clergymen in the highest age group, with a high specific death rate, is much greater than the proportion of railwaymen in the same group. The crude death rate for clergymen is therefore relatively too high.)

References

1. Irving Fisher, *The Making of Index Numbers* (Houghton Mifflin Co., 1923).
2. Bruce D. Mudgett, *Index Numbers* (John Wiley & Sons, Inc., 1951).
3. Raymond Pearl, *Medical Biometry and Statistics* (W. B. Saunders Co., 1930).

576
575
574
573
572
571
570
569
568
567

61

62



CHAPTER VI

STANDARD DEVIATION AND OTHER MEASURES OF DISPERSION

6.1 Various Measures of Dispersion. As already mentioned, the dispersion of a distribution is indicated by the extent to which observed values of the variate tend to spread over an interval rather than to cluster closely around a central average. One measure of dispersion is the quartile deviation, described in §3.4, but there are also in common use the *range*, the *mean absolute deviation*, and the *standard deviation*. Of these, the standard deviation is by far the most important.

6.2 The Range. The range is the most simple and obvious measure of dispersion. It is the length of an interval which just covers the highest and the lowest observed values in a set, and thus measures the spread in the most direct way possible. Among the grades of Table 2, §1.8, the highest is 98 and the lowest 34, so that the range is $98 - 34$, or 64.

If the observed values are *measurements* which are recorded with a finite "step" (for example, lengths recorded to the nearest half inch), the range as calculated by subtracting the lowest value from the highest should, strictly speaking, be increased by this step. Thus, if the numbers in Table 2 represented the results of 100 measurements of lengths, in inches to the nearest half inch, the number 98 might mean a length as great as 98.25 in. and the number 34 a length as small as 33.75 in. The range would therefore be 64.5 in.

In grouped frequency distributions the range is usually taken as the difference between the highest and lowest class marks, but if the total frequency is sufficiently great for even the end classes to contain several members, it is probably better to subtract the lower boundary of the first class from the upper boundary of the last class. The range, however, is not a very precise measure, being greatly affected by the presence of a few exceptionally large or exceptionally small values among those observed. Thus in a sample of 738 Welshmen the range of weight was 190 lb, but this range was reduced to 120 lb merely by omitting the 5 heaviest men. (See §0.4, Reference 22, p. 111.)

The outstanding merit of the range is its simplicity, and for this reason it is used a great deal in industrial *quality control*. In many modern factories a running check is kept on the quality of the output by taking regular samples, and noting both the mean (as a measure of the average level) and the range (as a measure of variability). The variable measured may be, for example, the breaking strength of a sample of cloth or the width of a slot in a machine part. For practical reasons the sample must be small (five is a common size),

and the computations must be within the capabilities of a works foreman, so that he can take action at once if action seems to be called for. In routine control all he needs to do is to add five readings and subtract the lowest from the highest, plotting the results on prepared charts, and as long as the plotted points stay well inside the limits marked on the charts, the factory process may be assumed to be satisfactorily "in control."

The defects of the range are that it depends only on *two* values out of the whole set, that it is comparatively sensitive to fluctuations of sampling (different samples of the same size from the same population may have widely different ranges), and that it is difficult to work with mathematically. But in spite of this difficulty, fairly good tables now exist from which it can readily be determined whether an observed range in a sample of given size (at least from certain types of parent population) is, or is not, unusual. See §13.15.

6.3 Deviations. The deviation of a value x , from a fixed arbitrary value x_0 was defined in §4.5 as the difference $x - x_0$. In practice x_0 is almost always taken as the arithmetic mean \bar{x} , and if so we have the important result:

Theorem 1. *The arithmetic mean of the deviations is zero.*

Proof: Let $x_i' = x_i - \bar{x}$, and let the frequency corresponding to x_i be f_i . Then

$$\begin{aligned}\sum_{i=1}^k f_i x_i' &= \sum f_i x_i - \sum f_i \bar{x} \\ &= \sum f_i x_i - \bar{x} \sum f_i\end{aligned}$$

But by Eq. (4.2) $\sum f_i x_i = N\bar{x}$ and $\sum f_i = N$, whence $\sum f_i x_i' = 0$. Dividing by N , we have the result stated.

The greater the spread of a distribution, the greater numerically will be the largest deviations, and it seems reasonable to take, as a measure of the variation about the mean, a suitable average of *all* the deviations. The arithmetic mean of the deviations, as we have just seen, is zero, but we can take the arithmetic mean of the *absolute* deviations (disregarding signs), or the root mean square average (in which all the deviations are squared). The former procedure gives the *mean absolute deviation* and the latter the *standard deviation*.

6.4 Mean Absolute Deviation. The absolute value of any real number y , denoted by $|y|$, is the numerical value of y , with positive sign. That is, $|y| = y$ if $y \geq 0$, and $|y| = -y$ if $y < 0$. The *mean absolute deviation* (M.A.D.), often inaccurately called the *mean deviation*, is defined by

$$(6.1) \quad \text{M.A.D.} = \frac{1}{N} \sum f_i |x_i - \bar{x}|$$

Example. Find the mean absolute deviation for the grades in Table 14,§4.5, where the mean value of x is 72.5.

x	f	$ x - \bar{x} $	$f x - \bar{x} $
34.5	2	38	76
44.5	3	28	84
54.5	11	18	198
64.5	20	8	160
74.5	32	2	64
84.5	25	12	300
94.5	7	22	154
Total	100		1036

$$\text{M.A.D.} = \frac{1036}{100} = 10.36$$

There is some theoretical reason for defining the M.A.D. in terms of deviations from the *median*. It can be proved that $\sum f_i |x_i - x_0|$ is a minimum, for any choice of x_0 , when x_0 is the median. However, the use of the mean is conventional.

For a grouped distribution a *coded* method of calculating the M.A.D. can be used, but the formula is not very convenient to use. The mean absolute deviation is generally unwieldy in mathematical discussions, and its chief use in statistics is in situations where occasional large and erratic deviations are likely to occur. The standard deviation, which uses the squares of these large deviations, tends to overemphasize them.

6.5 The Standard Deviation. The common measure of dispersion, to be preferred in most circumstances, is the root-mean-square average of the deviations from the mean. The name *standard deviation* for this quantity was proposed by Karl Pearson. We shall denote it by

$$(6.2) \quad s_x = \left[\frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 \right]^{1/2}, \quad N = \sum_i f_i$$

The square root has the effect of making the unit of s_x the same as the unit of x_i . If x_i is in feet, $(x_i - \bar{x})^2$ is in square feet, and s_x again in feet. The quantity s_x^2 is called the *variance*, and in many ways it is more fundamental than the standard deviation.

The traditional symbol for the standard deviation is the Greek letter σ (sigma). Many writers even use such phrases as "a deviation of three sigmas" meaning a deviation three times as great as the standard deviation. In recent years, however, the practice has been to distinguish between a *sample* and the *population* from which it is drawn, by using Latin letters for quantities characterizing the sample and Greek letters for corresponding

quantities characterizing the population. With this convention (which we shall find very useful in the mathematical discussions of Part Two) the variance of a population is denoted by σ_x^2 and that of a sample from the population by s_x^2 . One of the main problems of statistics is that of estimating characteristics of a population from the characteristics of a finite, and possibly small, sample, but we can rarely calculate σ_x^2 directly. We are therefore usually concerned in an actual calculation with s_x^2 .

It is proved in Part Two (§7.9) that if s_x^2 is the variance of a sample of N , the "best" estimate of the population variance σ_x^2 (best in a certain sense which is explained there*) is not s_x^2 but $Ns_x^2/(N-1)$. If, however, we define s_x by

$$(6.3) \quad s_x = \left[\frac{1}{N-1} \sum f_i (x_i - \bar{x})^2 \right]^{1/2}$$

it will be true that, in the same sense as before, s_x^2 itself is the best estimate of σ_x^2 . For this reason several authorities (among others, R. A. Fisher and S. S. Wilks) prefer to use the definition (6.3) instead of (6.2). Since the justification for using (6.3) belongs to more advanced mathematical statistics, and since its use prevents the definition of s_x as an average of deviations, we shall continue to use the form (6.2). Of course, for large values of N the two forms are practically indistinguishable.

6.6 Calculation of the Standard Deviation. Equation (6.2), although the definition of s_x , is not well suited for actual computation. The quantities $x_i - \bar{x}$ are usually awkward numbers to square, and if they are not carried out to several figures an appreciable error may be introduced into the result.

From (6.2), we have

$$\begin{aligned} Ns_x^2 &= \sum f_i (x_i - \bar{x})^2 \\ &= \sum f_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \sum f_i x_i^2 - 2\bar{x} \sum f_i x_i + \bar{x}^2 \sum f_i \end{aligned}$$

But $\sum f_i x_i = N\bar{x}$ and $\sum f_i = N$ so that we obtain, on substituting for these sums,

$$\begin{aligned} Ns_x^2 &= \sum f_i x_i^2 - 2N\bar{x}^2 + N\bar{x}^2 \\ &= \sum f_i x_i^2 - N\bar{x}^2 \\ (6.4) \quad &= \sum f_i x_i^2 - \frac{1}{N} (\sum f_i x_i)^2 \end{aligned}$$

This is the usual form for the calculation of s_x for a discrete variate, or from a set of *ungrouped* values of x . If N is fairly small, all the x_i may be treated individually, in which case all the f_i will be equal to 1 and we shall have

$$(6.5) \quad Ns_x^2 = \sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2$$

* See also §7.9 of this book.

The arithmetic can often be simplified by subtracting a suitable fixed value x_0 from each x_i . If $u_i = x_i - x_0$, and consequently, by (4.4), $\bar{u} = \bar{x} - x_0$,

$$\begin{aligned} Ns_x^2 &= \sum f_i [(x_i - x_0) - (\bar{x} - x_0)]^2 \\ &= \sum f_i (u_i - \bar{u})^2 \\ (6.6) \qquad &= \sum f_i u_i^2 - \frac{1}{N} (\sum f_i u_i)^2 \end{aligned}$$

The calculation can therefore be carried out entirely in terms of the variate u , and no adjustment is needed in the result.

Example 1. Find the variance of the integers 1 to 10 inclusive. Here we use (6.5). By Theorems 4 and 5 of §4.2,

$$\sum_1^{10} x_i = 10(10 + 1)/2 = 55$$

and

$$\sum_1^{10} x_i^2 = 10(10 + 1)(20 + 1)/6 = 385$$

so that

$$\begin{aligned} 10s_x^2 &= 385 - (55)^2/10 \\ &= 82.5 \end{aligned}$$

and

$$s_x^2 = 8.25$$

TABLE 22. AVERAGE YIELDS OF CORN IN BUSHELS PER ACRE
FOR A CERTAIN SECTION IN ILLINOIS FROM 1901-1920

Year	Yield (x)	u	u^2
1901	21	-15	225
1902	39	3	9
1903	32	-4	16
1904	37	1	1
1905	40	4	16
1906	36	0	0
1907	36	0	0
1908	32	-4	16
1909	36	0	0
1910	39	3	9
1911	33	-3	9
1912	40	4	16
1913	27	-9	81
1914	29	-7	49
1915	36	0	0
1916	30	-6	36
1917	38	2	4
1918	36	0	0
1919	36	0	0
1920	35	-1	1
Totals	$N = 20$	-32	488

Example 2. Find the mean yield of corn and the standard deviation of yields, for the years 1901–1920, from the data of Table 22.

Here the items are ungrouped and can be treated individually. (It is hardly worth while to rewrite the table merely for the sake of grouping together the six values of x each equal to 36.) Subtracting 36 from each x , we obtain the values of u shown.

$$\sum u_i = -32 \quad \text{and} \quad \sum u_i^2 = 488$$

so that

$$20s_x^2 = 488 - (32)^2/20 = 436.8$$

giving $s_x = 4.67$.

Since $\bar{u} = -32/20 = -1.6$, we have $\bar{x} = 36 - 1.6 = 34.4$. The mean and standard deviation of the annual yields are therefore 34.4 bushels and 4.67 bushels, per acre.

Example 3. Calculate the standard deviation of the number of sixes in a throw of 12 dice (Table 10, §2.3).

There is no need to change the variable since the x , are small numbers and we can use (6.4). The calculations are shown in Table 23. Column 4 is easily obtained by multiplying together the columns for x and fx .

TABLE 23. NUMBER OF SIXES IN A THROW OF 12 DICE

x	f	fx	fx^2	
0	447	0	0	
1	1145	1145	1145	
2	1181	2362	4724	$\bar{x} = 8191/4096 = 2.000$
3	796	2388	7164	
4	380	1520	6080	$s_x^2 = \frac{23259}{4096} - \left(\frac{8191}{4096}\right)^2$
5	115	575	2875	$= 1.679$
6	24	144	864	
7	7	49	343	$s_x = 1.296$
8	1	8	64	
Total	4096	8191	23,259	

6.7 Standard Deviation of a Grouped Continuous Variate. The method of §4.5 for calculating the arithmetic mean of a continuous variate grouped in classes may be extended to the calculation of the variance. If c is the class interval and x_0 the class mark for one of the central classes of the distribution, the coded variate u is given by $x_i = cu_i + x_0$, and the u_i are simple consecutive integers when the classes are all equal in width.

By (4.4), $\bar{x} = c\bar{u} + x_0$, so that

$$x_i - \bar{x} = c(u_i - \bar{u})$$

Hence,

$$\begin{aligned} \sum f_i(x_i - \bar{x})^2 &= c^2 \sum f_i(u_i - \bar{u})^2 \\ &= c^2 \left[\sum f_i u_i^2 - \frac{1}{N} \left(\sum f_i u_i \right)^2 \right] \\ &= c^2 N s_u^2 \end{aligned}$$

where s_u^2 is the variance of the u variate, as defined in (6.4). We have therefore,

$$(6.7) \quad s_z^2 = c^2 s_u^2$$

so that we merely have to multiply the calculated s_u by c to get the value of s_z .

Example 4. Calculate the standard deviation for the grades of Table 14, §4.5.

Here it is merely necessary to extend the table so as to include another column for fu^2 , found by multiplying together u and fu . The computations are shown in Table 24. The last column is a check column (§6.8).

TABLE 24. GRADES OF 100 STUDENTS (TABLE 2)

Class Mark x	Frequency f	u	fu	fu^2	$f(u+1)^2$
34.5	2	-4	-8	32	18
44.5	3	-3	-9	27	12
54.5	11	-2	-22	44	11
64.5	20	-1	-20	20	0
74.5	32	0	0	0	32
84.5	25	1	25	25	100
94.5	7	2	14	28	63
Total	100		-20	176	236

$$\bar{u} = -0.2$$

$$100s_u^2 = 176 - 400/100 = 172$$

$$s_u = (1.72)^{1/2} = 1.31$$

$$s_z = cs_u = 13.1$$

6.8 Charlier Check. In all but the simplest statistical calculations, and particularly in the more long and complicated ones, it is desirable to have systematic checks on the arithmetic. One such check, due to L. V. Charlier, may conveniently be introduced into the work-sheet for the standard deviation. The check involves forming a column of values of $f(u+1)^2$, and depends on the identity:

$$(6.8) \quad \begin{aligned} \sum f(u+1)^2 &= \sum f(u^2 + 2u + 1) \\ &= \sum fu^2 + 2\sum fu + \sum f \end{aligned}$$

To form the product $f(u+1)^2$ we multiply each value of f by the value of u in the next line below, and multiply the result again by the same u . In Table 24, $\sum f(u+1)^2 = 236$, and the right-hand side of (6.8) gives $176 + 2(-20) + 100$, which is also 236, so that the arithmetic is checked.

6.9 Grouping Error of the Standard Deviation. When dealing with a grouped variate we calculate the mean and standard deviation as though all

the values within a class were equal to the class mark. In the calculation of the mean for a fairly large sample this assumption usually introduces no appreciable error, because the errors introduced in the region of values below the mean tend to be compensated by opposite errors in the region above the mean. When the values are squared, however, the compensation of errors is far from exact, particularly with a coarse grouping. To give a simple and artificial illustration, suppose the population of the following table is grouped in two classes, thus:

x	1	2	3	4	5	6	x	2	5
f	10	20	30	30	20	10	f	60	60
<i>Ungrouped</i>							<i>Grouped</i>		

It is easily verified that $\sum fx$ is 420 in both cases, but $\sum fx^2$ is increased from 1700 to 1740, so that s_x^2 is increased from 1.92 to 2.25.

Unless we have the original data and work out the standard deviation exactly, we cannot tell how large this grouping error really is. However, for samples of a few hundred which are of the hump-backed type, tailing off gradually at both ends, it has been shown that, *on the whole*, an improved estimate of the true variance is given by subtracting $c^2/12$ from the variance as calculated from the grouped distribution. This correction is known as *Sheppard's correction*. It is most easily applied by subtracting $1/12$ ($= 0.0833$) from the calculated value of s_u^2 .

As applied to Table 24, this correction would reduce s_u^2 from 1.72 to 1.6367, giving $s_u = 1.28$ and $s_x = 12.8$. The true value can be calculated from the data in Table 2, by using formula (6.5), and is 12.9, so that here the correction does improve the result. However, the total frequency is rather small in this example for the application of Sheppard's correction. With small samples the error due to the fluctuation from one random sample to another is much larger than the correction, so that the latter is an unnecessary refinement.

6.10 Meaning of the Standard Deviation. Because of the comparatively elaborate calculation required to determine it, the standard deviation is not as easy to visualize as the range, or even as the quartile deviation. Considered as an *average* deviation, it is not as simple an average as the mean absolute deviation. Yet it is so important a concept that every effort should be made to get a clear idea of its meaning and of its approximate size in relation to the other measures of dispersion.

Many actually occurring distributions have frequency curves of the uni-modal type, with more or less flat "tails" at both ends, and are roughly symmetrical. Such frequency curves can be regarded as approximations to a well-defined mathematical curve known as the "normal curve," which will be discussed in a later chapter. For the normal curve the range is infinite, but 99.7% of the whole area under the curve is contained within an interval

from 3σ below the mean to 3σ above the mean (σ being the standard deviation). For a sample of two or three hundred individuals, therefore, the effective range is about six times the standard deviation, and this fact is a useful rough guide to the expected magnitude of the standard deviation. For very large samples the range is greater than this, and for small samples less, relative to the standard deviation.

Another property of the normal curve is that the area between ordinates at distances σ above and below the mean is 0.683 times the total area. Roughly, for any distribution not too different from the "normal" type, about two-thirds of the whole distribution is contained in an interval from $\bar{x} - s_x$ to $\bar{x} + s_x$. This is perhaps the clearest picture one can get of the standard deviation, although it does not hold for very skew distributions. It is illustrated in Fig. 21 for the distribution of Table 24. It may be noted that for this distribution the interval is from 59.7 to 85.3. The number of grades in Table 2 (§1.8) actually within these limits is 68, which is just about what we should expect. Since one-half, exactly, of the distribution lies between the quartiles, the standard deviation is larger than the quartile deviation. The ratio s_x/Q is usually close to $1\frac{1}{2}$.

For a normal distribution the ratio of the standard deviation to the mean absolute deviation is 1.253, or almost $1\frac{1}{4}$. The mean absolute deviation for the grades in Table 24 is 10.36, so that we should expect a standard deviation of about 13, as we actually find.

We have already mentioned in §4.8 that when a distribution is represented by a histogram, the centroid (or center of gravity) lies on the ordinate through the arithmetic mean. If we think of the histogram as cut out of a thin uniform metal plate and rotated about an axis in its plane through the centroid, it will have a certain "moment of inertia." The distance from the axis at which the entire mass could be considered as concentrated without changing the moment of inertia is called the *radius of gyration* and is equal to the standard deviation of the distribution. This illustration may help the mechanically minded student.

6.11 Relative Dispersion. The size of observed values usually influences not only the mean but also deviations from the mean. In other words, the magnitudes of the deviations from the mean seem to be dependent, in some degree, upon the magnitude of the mean. In comparing dispersion in dis-

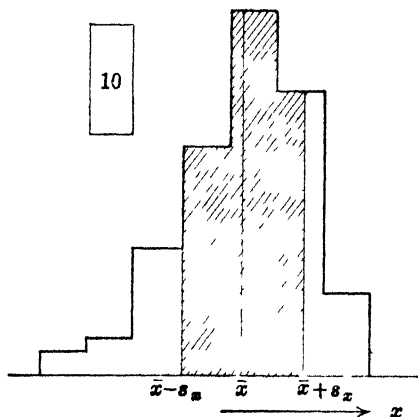


FIG 21 STANDARD DEVIATION

tributions, we may correct for differences in the average magnitudes of positive values by taking the ratio of the standard deviation to the mean. Thus, the quantity

$$(6.9) \quad V = \frac{s_x}{\bar{x}}$$

is known as the *coefficient of variation*. It is obviously an abstract number, being independent of the units of measurement, and it is usually expressed as a percentage.

The use of (6.9) may be misleading in situations where the origin from which the data are measured is somewhat arbitrary. Cases in point are temperature measurements and certain psychological data.

6.12 Some Theorems on Variance. One of the most important properties of the standard deviation (or of the variance, which is its square) is the comparative ease with which it can be manipulated in the theoretical discussions of mathematical statistics. Some illustrations of this property are given in the present section.

Theorem 1. *The sum of squares of deviations of the variate values from their mean is less than the sum of squares of deviations from any other number. That is, for an arbitrary x_0 , not equal to \bar{x} ,*

$$(6.10) \quad Ns_x^2 < \sum_{i=1}^N (x_i - x_0)^2$$

(For convenience the x_i are treated individually. There may, of course, be f_i values all equal to x_i .)

$$\text{Proof: } x_i - x_0 = x_i - \bar{x} + (\bar{x} - x_0)$$

Therefore

$$\begin{aligned} \sum (x_i - x_0)^2 &= \sum (x_i - \bar{x})^2 + \sum (\bar{x} - x_0)^2 + 2\sum (x_i - \bar{x})(\bar{x} - x_0) \\ &= \sum (x_i - \bar{x})^2 + N(\bar{x} - x_0)^2 + 2(\bar{x} - x_0)\sum (x_i - \bar{x}) \end{aligned}$$

But $\sum (x_i - \bar{x}) = N\bar{x} - N\bar{x} = 0$, and $\sum (x_i - \bar{x})^2 = Ns_x^2$. Hence

$$(6.11) \quad \sum (x_i - x_0)^2 = Ns_x^2 + N(\bar{x} - x_0)^2$$

Since the last term on the right-hand side is necessarily positive, whether x_0 is greater or less than \bar{x} , the theorem follows.

Theorem 2. *Let there be one set of n_1 variates x_{1i} ($i = 1, 2, \dots, n_1$) and another set of n_2 variates x_{2i} ($i = 1, 2, \dots, n_2$) and let \bar{x} be the mean of the combined sets (Theorem 10, §4.6). The variance s^2 of the set formed by the com-*

bination of these two sets is given by the following formula:

$$(6.12) \quad Ns^2 = \sum_1^{n_1} (x_{1i} - \bar{x})^2 + \sum_1^{n_2} (x_{2i} - \bar{x})^2$$

where

$$N = n_1 + n_2$$

Proof: The proof consists in showing that

$$\sum_1^{n_1} (x_{1i} - \bar{x})^2 + \sum_1^{n_2} (x_{2i} - \bar{x})^2 = \sum_1^{n_1+n_2} (x_i - \bar{x})^2$$

which is left as an exercise for the student.

The foregoing theorem is not very important in itself, but it is useful in proving the next theorem which gives the relation between the variance of a composite set and the variances of subsets.

Theorem 3. *Let the frequency, mean, and standard deviation be denoted by n_1, \bar{x}_1 , and s_1 for one set of variates and by n_2, \bar{x}_2 , and s_2 for a second set. The variance s^2 of the composite set is given by the following relation:*

$$Ns^2 = n_1s_1^2 + n_2s_2^2 + n_1d_1^2 + n_2d_2^2$$

where $N = n_1 + n_2$, $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$, and \bar{x} is the mean of the composite set.

Proof: For the n_1 set, \bar{x} may be regarded as an arbitrary point x_0 . Hence by (6.11) we have, after rearranging terms,

$$\frac{1}{n_1} \sum_1^{n_1} (x_{1i} - \bar{x}_1)^2 = \frac{1}{n_1} \sum_1^{n_1} (x_{1i} - \bar{x})^2 - (\bar{x}_1 - \bar{x})^2$$

Multiplying through by n_1 this becomes

$$(6.13) \quad n_1s_1^2 = \sum_1^{n_1} (x_{1i} - \bar{x})^2 - n_1d_1^2$$

Similarly for the n_2 group we have

$$(6.14) \quad n_2s_2^2 = \sum_1^{n_2} (x_{2i} - \bar{x})^2 - n_2d_2^2$$

Adding (6.13) and (6.14), and using (6.12), we obtain

$$n_1s_1^2 + n_2s_2^2 = Ns^2 - n_1d_1^2 - n_2d_2^2$$

Hence

$$(6.15) \quad Ns^2 = n_1s_1^2 + n_2s_2^2 + n_1d_1^2 + n_2d_2^2$$

For k sets combined into a single set we can generalize (6.15) into the following relation:

$$(6.16) \quad Ns^2 = \sum_1^k n_i s_i^2 + \sum_1^k n_i d_i^2$$

where $N = \sum_1^k n_i$ and $d_i = \bar{x}_i - \bar{x}$. It is interesting to observe that $\frac{1}{N} \sum_1^k n_i d_i^2$ is the variance of the means of the subsets. Thus we have the important relation

$$(6.17) \quad s^2 = \frac{1}{N} \sum_1^k n_i s_i^2 + s_z^2$$

which shows that the total variance may be broken up into two parts, one of which is the weighted mean of the variances in the subsets and the other is the variance of their means. These two parts are sometimes called the average variance *within* classes and the variance *between* the means of the classes. They become very important in the "Analysis of Variance." (See §12.16 and also Part Two, Chapter IX.)

Corollary. Equation (6.15) may also be written

$$(6.18) \quad Ns^2 = n_1 s_1^2 + n_2 s_2^2 + \frac{n_1 n_2}{N} (\bar{x}_1 - \bar{x}_2)^2$$

$$\begin{aligned} \text{Proof:} \quad n_1 d_1^2 + n_2 d_2^2 &= n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2 \\ &= n_1 \bar{x}_1^2 + n_2 \bar{x}_2^2 - 2\bar{x}(n_1 \bar{x}_1 + n_2 \bar{x}_2) + (n_1 + n_2) \bar{x}^2 \end{aligned}$$

By Theorem 10, §4.6, $n_1 \bar{x}_1 + n_2 \bar{x}_2 = N\bar{x}$, and also $n_1 + n_2 = N$. Therefore

$$\begin{aligned} (6.19) \quad n_1 d_1^2 + n_2 d_2^2 &= n_1 \bar{x}_1^2 + n_2 \bar{x}_2^2 - N\bar{x}^2 \\ &= n_1 \bar{x}_1^2 + n_2 \bar{x}_2^2 - (n_1 \bar{x}_1 + n_2 \bar{x}_2)^2 / N \end{aligned}$$

The coefficient of \bar{x}_1^2 is $n_1 - \frac{n_1^2}{n_1 + n_2} = \frac{n_1 n_2}{n_1 + n_2} = \frac{n_1 n_2}{N}$ and similarly for \bar{x}_2^2 .

The coefficient of $\bar{x}_1 \bar{x}_2$ is $-2n_1 n_2 / N$. Hence

$$\begin{aligned} n_1 d_1^2 + n_2 d_2^2 &= \frac{n_1 n_2}{N} (\bar{x}_1^2 + \bar{x}_2^2 - 2\bar{x}_1 \bar{x}_2) \\ &= \frac{n_1 n_2}{N} (\bar{x}_1 - \bar{x}_2)^2 \end{aligned}$$

The equation (6.18) then follows from (6.15). This form cannot be generalized to k sets.

Exercises

1. What is the range in (a) the weights of Table 12, §2.5; (b) the scores in Exercise 9 below?

2. (*E. S. Pearson*). The following data represent the percentage of ash-content in 280 wagon tests of a certain kind of coal. Find the mean and the standard deviation of the distribution:

<i>Percentage Ash-Content</i>	<i>Frequency</i>
3.0- 3.9	1
4.0- 4.9	7
5.0- 5.9	28
6.0- 6.9	78
7.0- 7.9	84
8.0- 8.9	45
9.0- 9.9	28
10.0-10.9	7
11.0-11.9	2

Ans. $\bar{x} = 7.35\%$, $s_x = 1.36\%$.

3. (*Camp*). Find the mean wage and the standard deviation of wages for the following distribution:

<i>Class</i>	<i>Frequency</i>
\$4.50- 5.99	43
6.00- 7.49	99
7.50- 8.99	152
9.00-10.49	178
10.50-11.99	160
12.00-13.49	40
13.50-14.99	25
15.00-16.49	3

Ans. $N = 700$, $\bar{x} = \$9.42$, $s_x = \$2.19$.

4. Compute the mean absolute deviation for the data of Exercise 2. Find the ratio of the M.A.D. to the standard deviation.

5. Find the mean and standard deviation for the data of Table 11, §2.3. Compare the range with the standard deviation. *Ans.* $\bar{x} = 6.139$, $s_x = 1.712$, range = 10.

6. Find the mean and standard deviation for the distribution of lengths of telephone calls given in Table 25. Use the Charlier check and Sheppard's correction.

TABLE 25. DISTRIBUTION OF LENGTHS OF
995 TELEPHONE CALLS. TIME IN SECONDS

<i>Time</i>	<i>Number of Calls</i>
0-99	1
100-199	28
200-299	88
300-399	180
400-499	247
500-599	260
600-699	133
700-799	42
800-899	11
900-999	5

7. Calculate the quartile deviation for the data of Table 25, and find the ratio to the standard deviation. *Ans.* 0.69.

8. Find the percentage of values in the distribution of Table 25 outside the limits $\bar{x} \pm s_x$, $\bar{x} \pm 2s_x$ and $\bar{x} \pm 3s_x$, respectively. *Ans.* 32.7, 5.0, 0.4.

Hint. Assume that the values within any class are equally spaced throughout the class interval. It will be helpful to construct a histogram and mark off the limits mentioned with ordinates. The calculation is similar to that for percentile ranks.

9. Find the mean and standard deviation for the following set of 25 scores: 82, 86, 75, 78, 72, 79, 63, 65, 67, 75, 68, 70, 79, 78, 51, 58, 65, 69, 68, 83, 80, 42, 43, 48, 47. *Ans.* $\bar{x} = 67.6$, $s_x = 12.7$.

10. The following data were obtained in a mental test for 290 prospective employees: mean = 43.33 pts., $s_x = 9.25$ pts. The percentage of standard production attained by these same persons after being employed varied about a mean of 92.02% with a standard deviation of 24.47%. Compare the relative dispersions in mental test and production ability.

11. Find \bar{x} , s_x , and the M.A.D. for the following distribution:

x	2	4	6	8	10
f	1	4	6	4	1

12. Show that (6.5) may be written

$$s_x = [N \sum x^2 - (\sum x)^2]^{1/2} / N$$

(probably the most convenient formula for machine computation with ungrouped data).

13. For a set of ungrouped values the following sums are found:

$$N = 15, \quad \sum x = 480, \quad \sum x^2 = 15,735$$

Find the mean and the standard deviation.

14. If only two values, x_1 and x_2 , are obtained for the variate x , and if $\bar{x} = (x_1 + x_2)/2$, show that

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 = (x_1 - x_2)^2 / 2$$

and that consequently

$$s_x = |x_1 - x_2| / 2.$$

15. Verify the identity

$$3(x_1 - x_2)^2 + (x_1 + x_2 - 2x_3)^2 = 6 \sum_{i=1}^3 (x_i - \bar{x})^2$$

and so obtain an expression for the variance of three values x_1 , x_2 and x_3 , with mean \bar{x} .

16. Given the following information about two sets of data:

$$\text{I} \quad n_1 = 20, \quad \bar{x}_1 = 25, \quad s_1^2 = 5$$

$$\text{II} \quad n_2 = 30, \quad \bar{x}_2 = 20, \quad s_2^2 = 4$$

Find the mean and variance of the composite set.

17. For a group of 50 boys the mean score and standard deviation of scores on a test are 59.5 and 8.38. For a group of 40 girls the mean and standard deviation are 54.0 and 8.23 on the same test. Find the mean and standard deviation for the combined group of 90 children.

18. Calculate the mean and standard deviation of the first 25 positive integers

19. Prove that the variance of the first N positive integers is $(N^2 - 1)/12$.

20. For eight related distributions the following values are obtained:

Distribution	1	2	3	4	5	6	7	8
Frequency	7	14	32	49	55	54	35	14
Mean	67.86	72.14	81.87	84.80	85.73	90.92	95.57	105.0
Variance	106.1	191.8	246.5	283.6	257.6	294.5	222.5	71.43

Find the mean and variance of the whole distribution formed by the combination of these eight. *Ans.* 87.31, 303.1.

Hint. Use equations (4.8) and (6.16).

21. Show that (4.8) and (6.15) may be written

$$\bar{x}_1 = [N\bar{x} - \sum_{i=2}^k n_i \bar{x}_i] / n_1$$

$$s_1^2 = [N(s^2 + \bar{x}^2) - n_2(s_2^2 + \bar{x}_2^2)] / n_1 - \bar{x}_1^2$$

(These forms are required in the next exercise.)

22. In a certain distribution of $N = 25$ measurements, it was found that $\bar{x} = 56$ inches and $s = 2$ inches. After these results were computed it was discovered that a mistake had been made in one of the measurements which was recorded as 64 inches. Find the mean and standard deviation if the incorrect variate, 64, is omitted.

Hint. Let $n_1 = 24$, $n_2 = 1$. Then $\bar{x}_2 = 64$ and $s_2 = 0$. To find \bar{x}_1 and s_1 use formulas in Exercise 21 above.

23. If two or more variates are deleted from a distribution for which N , \bar{x} , and s are given, show how to compute the mean and variance of the remaining variates.

24. Consider a composite set consisting of k subsets and let s_i^2 and n_i denote, respectively, the variance and number of variates in the i th subset, and $N = \sum_1^k n_i$.

(a) If the subsets have equal means, show that the variance of the composite set is given by

$$s^2 = \frac{1}{N} \sum_1^k n_i s_i^2$$

(b) If the subsets each contain the same number of variates and have equal means, show that

$$s^2 = \frac{1}{k} \sum_1^k s_i^2$$

CHAPTER VII

MOMENTS. SKEWNESS AND KURTOSIS

7.1 Populations and Samples. One of the general problems of statistics is to summarize and characterize data. In the words of R. A. Fisher,

“A quantity of data which by its mere bulk may be incapable of entering the mind is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.”*

Among these “relatively few quantities” are those which are known as *moments*. Two of them, the arithmetic mean and the variance, have already been discussed and two others are in fairly common use, but a whole series of moments can be defined. The higher moments are principally used to characterize populations rather than samples, and therefore it will be well to clarify the distinction between a sample and the population from which it is drawn.

There are three kinds of population. The first includes finite and actually existing populations which, although large, can be enumerated if necessary. Examples are the total number of persons living in the United States or the total number of peach farms in the state of Georgia. Since complete censuses are time-consuming and costly, it is usual to obtain information about such populations by investigating samples which are more or less randomly chosen.

The second kind of population is a generalization from experience and is indefinitely large, such as, for instance, the total number of throws that might conceivably be made in unlimited time with a particular pair of dice. Any actual set of throws, however numerous, can be regarded as a sample from this practically infinite population. A third kind is a purely hypothetical population which can be completely described mathematically. That is, the distribution of values in the population is given by a mathematical formula. This type of population is used in the process known as *curve-fitting* or *graduation*, in which an actually occurring distribution is replaced for the purposes of further discussion by a mathematically described distribution which seems to have similar characteristics. If the “fit” is satisfactory, as judged by a test which we shall describe later, we can regard the observed sample as coming from a population which has the characteristics of the mathematical

* See Reference 1.

distribution. In the most common method of curve-fitting, moments play an important part.

7.2 Moments about the Origin. The variate x with which we are concerned may be discrete or continuous. If discrete, x may take values x_1, x_2, \dots with frequencies f_1, f_2, \dots and a total population frequency M ; if continuous, x_1, x_2, \dots are the class marks of classes with corresponding frequencies. When the population is indefinitely large, it does not make sense to speak of the frequencies f_i , but we can still speak of the *proportion* of values p_i lying in the i th class. Similarly, as we shall see in the next chapter, when the population is described by a continuous mathematical function we can talk of the proportion lying between any two given values of x , x_1 and x_2 . For a finite population, $p_i = f_i/M$ and when M is large this proportion can be regarded as the probability (see Chapter IX) that a value selected at random from the population will lie in the i th class. We shall define our moments first for a finite population. The r th moment about the origin of x is given by

$$(7.1) \quad \mu_r' = \frac{1}{M} \sum_i f_i x_i^r, \quad r = 0, 1, 2, 3, \dots$$

where $\sum f_i = M$.

For a practically infinite population we write $f_i = Mp_i$, and so obtain

$$(7.2) \quad \mu_r' = \sum_i p_i x_i^r$$

For $r = 0$, $x_i^0 = 1$, and so

$$(7.3) \quad \mu_0' = 1$$

For $r = 1$, we obtain

$$(7.4) \quad \mu_1' = \frac{1}{M} \sum_i f_i x_i \quad \text{or} \quad \sum_i p_i x_i$$

which is simply the arithmetic mean. For a population we shall denote the mean by μ .

For a sample of N , the moments about the origin may be defined by

$$(7.5) \quad m_r' = \frac{1}{N} \sum_i f_i x_i^r, \quad r = 0, 1, 2, \dots$$

where $\sum f_i = N$.

The first moment m_1' is then the arithmetic mean \bar{x} of the sample. Similarly $m_2' = \frac{1}{N} \sum f_i x_i^2$, the arithmetic mean of squares of the values x_i , and so for moments of higher order.

The term "moment" has its origin in mechanics where we speak of the "moment of a force." Suppose we have a rigid bar, called a lever, with one point of support known as a fulcrum (Fig. 22). If a force f_1 is applied to the lever at a distance x_1 from the fulcrum O , the product $x_1 f_1$ is called the moment of the force.

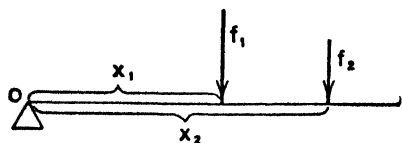


FIG. 22

If there are two or more such forces f_1, f_2, \dots, f_k , acting in the same direction, and at the distances x_1, x_2, \dots, x_k , respectively from O , the total moment of all these forces is

$$f_1 x_1 + f_2 x_2 + \dots + f_k x_k = \sum f_i x_i.$$

If the distances x are squared, we have $\sum f_i x_i^2$ as the total second moment, and $\sum f_i x_i^r$ represents the r th moment.

We have seen that in calculating the arithmetic mean and the standard deviation for a continuous variable, grouped in classes, it is convenient to change the variate from x to u , where

$$u = (x - x_0)/c$$

In the same way, in calculating moments of any order, it is often convenient to obtain first the moments in terms of u , namely,

$$(7.6) \quad m_{r,u'} = \frac{1}{N} \sum f_i u_i^r, \quad r = 0, 1, 2, \dots$$

Here we are using two subscripts on m , the first denoting the order of the moment and the second the variate. The second subscript can be omitted when there is no ambiguity.

7.3 Moments about the Mean. The most important set of moments in statistical theory is obtained by shifting the origin to the arithmetic mean. For a population of M , we have

$$(7.7) \quad \mu_r = \frac{1}{M} \sum_i f_i (x_i - \mu)^r, \quad r = 0, 1, 2, \dots$$

or, when M is infinite,

$$(7.8) \quad \mu_r = \sum_i p_i (x_i - \mu)^r$$

For $r = 0$ we have

$$(7.9) \quad \mu_0 = 1$$

For $r = 1$,

$$\begin{aligned}\mu_1 &= \frac{1}{M} \sum f_i(x_i - \mu) \\ &= \frac{1}{M} \sum f_i x_i - \mu \\ (7.10) \quad &= \mu - \mu = 0\end{aligned}$$

For $r = 2$,

$$\mu_2 = \frac{1}{M} \sum f_i(x_i - \mu)^2$$

which is the population variance and is commonly denoted by σ^2 .

For a sample of N , the corresponding moments are

$$(7.11) \quad m_r = \frac{1}{N} \sum f_i(x_i - \bar{x})^r$$

and here again $m_0 = 1$, $m_1 = 0$, and m_2 is the sample variance. Analogous formulas hold for the moments $m_{r,u}$ in terms of u .

If we think of weights proportional to the frequencies in each class suspended along a horizontal bar at the class marks, the bar will balance at its center of gravity which is at the point \bar{x} , equal to m_1' . Also if the bar is rotated about its center of gravity, its radius of gyration is the square root of the second moment $m_{2,x}$ and so is equal to the standard deviation of the distribution. There is no simple mechanical analogy for the higher moments.

7.4 Relations between the m_r' and the m_r . Even in the u variate, the m_r are troublesome to calculate directly from the definition, because \bar{u} will usually have to be taken out to several figures, and then $u - \bar{u}$ has to be squared, cubed, or raised to even higher powers. Therefore, instead of computing the m_r directly we first find the much simpler m_r' . By using the Binomial Theorem of algebra for a positive integral index, it is easy to obtain a series of expressions for m_r in terms of m_r' . If we work in the u variate (and drop the subscript u), we have, for example,

$$\begin{aligned}m_2 &= \frac{1}{N} \sum f_i(u_i - \bar{u})^2 \\ &= \frac{1}{N} \sum f_i u_i^2 - \frac{2}{N} \bar{u} \sum f_i u_i + \frac{\bar{u}^2}{N} \sum f_i \\ &= m_2' - 2\bar{u}m_1' + \bar{u}^2\end{aligned}$$

But $m_1' = \bar{u}$, so that

$$(7.12) \quad m_2 = m_2' - (m_1')^2$$

The student should be able to prove, by writing out the expanded forms of $(u_i - \bar{u})^3$ and $(u_i - \bar{u})^4$, that

$$(7.13) \quad m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3$$

and

$$(7.14) \quad m_4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 - 3(m_1')^4$$

Having obtained the moments in terms of u , it is merely necessary to multiply by the appropriate power of c to get the moments in terms of x . Thus,

$$\begin{aligned} m_{2,x} &= \frac{1}{N} \sum f_i (x_i - \bar{x})^2 \\ &= \frac{1}{N} \sum f_i (x_0 + cu_i - x_0 - c\bar{u})^2 \end{aligned}$$

by (4.3) and (4.4),

$$\begin{aligned} &= \frac{c^2}{N} \sum f_i (u_i - \bar{u})^2 \\ (7.15) \quad &= c^2 m_{2,u} \end{aligned}$$

It is easy to show similarly that

$$(7.16) \quad m_{r,x} = c^r m_{r,u}$$

These moments are therefore not affected by a change of origin, but only by a change of scale.

Similar relations hold for the population moments, for example,

$$(7.17) \quad \mu_2 = \mu_2' - (\mu_1')^2$$

$$(7.18) \quad \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$(7.19) \quad \mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

7.5 Calculation of Third and Fourth Moments. The method is a direct extension of that already used for the mean and the variance. The details of the calculation for the grades of Table 2 are set out in Table 26. The column fu^2 is found by multiplying fu by u , fu^3 by multiplying fu^2 by u , and so on. The last column is used for *Charlier's check*, which should be applied as soon as the other columns have been computed and added. The check depends on the identity:

$$(7.20) \quad \sum f(u+1)^4 = \sum fu^4 + 4\sum fu^3 + 6\sum fu^2 + 4\sum fu + \sum f$$

The value $u+1$ for any row except the last is the value of u in the next following row. The numbers are raised to the fourth power and multiplied by f . For the data in Table 26, $\sum f(u+1)^4 = 1220$, and the sum of terms

TABLE 26. MOMENTS FOR DISTRIBUTION OF GRADES

Data		Computations					
x	f	u	fu	fu^2	fu^3	fu^4	$f(u+1)^4$
34.5	2	-4	-8	32	-128	512	162
44.5	3	-3	-9	27	-81	243	48
54.5	11	-2	-22	44	-88	176	11
64.5	20	-1	-20	20	-20	20	0
74.5	32	0	0	0	0	0	32
84.5	25	1	25	25	25	25	400
94.5	7	2	14	28	56	112	567
Sums	100		-20	176	-236	1088	1220
$\frac{1}{N}$ Sums	1		-0.20 $m'_{1,u}$	1.76 $m'_{2,u}$	-2.36 $m'_{3,u}$	10.88 $m'_{4,u}$	For Charlier's check

on the right-hand side of (7.20) is $1088 + 4(-236) + 6(176) + 4(-20) + 100$, which is also 1220.

This check does not insure accuracy, because compensating errors might occur, but if the check is satisfied one feels confident to proceed. We divide the column sums by N to get the $m'_{r,u}$, and then compute the $m_{r,u}$ by (7.12), (7.13), and (7.14). Thus,

$$m_{2,u} = 1.76 - (-0.20)^2 = 1.72$$

$$m_{3,u} = -2.36 - 3(1.76)(-0.20) + 2(-0.20)^3 = -1.32$$

$$m_{4,u} = 10.88 - 4(-2.36)(-0.20) + 6(1.76)(-0.20)^2 - 3(-0.20)^4 = 9.41$$

As a check on the calculation we may compute $m'_{4,u}$ by the relation

$$\begin{aligned}
 m'_{4,u} &= \frac{1}{N} \sum f_i u_i^4 = \frac{1}{N} \sum f_i [(u_i - m'_1) + m'_1]^4 \\
 (7.21) \quad &= m_4 + 4m_3 m'_1 + 6m_2 (m'_1)^2 + (m'_1)^4
 \end{aligned}$$

(We have dropped the subscript u 's for convenience of printing.) This check can be handled readily on a computing machine.

7.6 Sheppard's Corrections for Grouping Errors. As explained in §6.9, the variance may be corrected (under suitable conditions) for the error caused by grouping. In the calculation of the higher moments the same assumption is made that all the values in any class may be taken as equal to the class

mark, and this assumption produces an appreciable error in the moments of even order. It may be shown that the corrections to the moments, in terms of the u variate, are

$$\text{corrected } m_2 = \text{uncorrected } m_2 - 1/12$$

$$\text{corrected } m_4 = \text{uncorrected } m_4 - \frac{1}{2} (\text{uncorrected } m_2) + 7/240$$

$$(1/12 = 0.08333, 7/240 = 0.02917)$$

The third moment needs no correction. As applied to Table 26,

$$\text{corrected } m_2 = 1.720 - 0.083 = 1.637$$

and

$$\text{corrected } m_4 = 9.410 - (1.720)/2 + 0.029 = 8.579$$

As remarked before, Sheppard's corrections are valid only for hump-backed ("bell-shaped") distributions with flat tails. They are not applicable to the J-shaped or U-shaped types. Moreover, they are a refinement which may not be consistent with the degree of accuracy in the original data or with the errors due to sampling fluctuation.

7.7 Standard Units. For the purposes of theoretical statistics it is often very convenient to express deviations from the mean in terms of the standard deviation as unit. We shall denote deviations so expressed by z , where

$$(7.22) \quad z = (x - \bar{x})/s_x$$

and say that they are *in standard units*, or *standardized*.

The significant characteristic of the z variate is its independence of the unit in which the original measurements were taken. For example, suppose we were concerned with a set of lengths. One distribution of variates would result if the measurements were made in feet. In this case x' , \bar{x} , and s_x would also be in feet. If the measurements were taken in inches, then x' , \bar{x} , and s_x would be in inches, and each of these values would be, numerically, twelve times as large as the corresponding numbers in the first distribution. However, the variates expressed in standard units would be the same for the two distributions. Thus if

$$\bar{x} = 50 \text{ ft} = 50(12) \text{ in}$$

and

$$s_x = 5 \text{ ft} = 5(12) \text{ in}$$

then for an individual measurement of $x = 60 \text{ ft} = 60(12) \text{ in}$, we have

$$z = \frac{10 \text{ ft}}{5 \text{ ft}} = \frac{10(12) \text{ in}}{5(12) \text{ in}} = 2$$

It is obvious, therefore, that standard units provide a basis for comparing distributions.

With the aid of a computing machine, a distribution may easily be transformed into standard units by the so-called continuous process. To illustrate, consider the data of Table 27, representing a distribution of weights

TABLE 27. STANDARD VALUES

x Class Mark (lb)	f	z
29.5	1	-3 154
33.5	14	-2.461
37.5	56	-1 768
41.5	172	-1.076
45.5	245	-0 383
49.5	263	0 310
53.5	156	1 002
57.5	67	1.695
61.5	23	2 388
65.5	3	3 081

(to the nearest pound) of 1000 8-year-old Glasgow schoolgirls. The mean and standard deviation are 47.712 lb and 5.774 lb, respectively, so that

$$z = \frac{x - 47.712}{5.774} = 0.17318x - 8.2627$$

Referring to the discussion of the continuous method given in the Introduction, we observe that here $k = -8.2627$, $n = 0.17318$, and we desire the values of z corresponding to the values of x given in Table 27. For the values of x such that $nx < k$, we write the above relation in the form

$$-z = 8.2627 - 0.17318x$$

The procedure* now is to register 8.262700 on the product register, punch the constant factor 0.17318 on the keyboard, and then by turning the crank backward so that the successive values of x appear on the revolution register, we subtract from k the products of this multiplier and the values of x . The various values of x are built over from one to another without clearing the dial. The resulting values of $-z$ are read at each stage from the product register until we get $-z = 0.383$. From here, $nx > k$, so we clear the dials and start over using the original form of the relation between x and z . We now

* If automatic machines are available the instructor will explain the procedure.

register -8.262700 on the product register by turning the crank backward, punch 0.17318 on the keyboard, and turn the crank forward to form the values of x on the revolution register. The values of z are read as before from the product register at each stage of the build-over process. In this way the set of standard values in Table 27 is obtained. We see from this table that a range of $z = \pm 3$ takes in practically all the values. This is typical of the more common distributions.

Some writers use X to denote the variate that we have called x , and use x to mean $X - \bar{X}$. In this notation $z = x/s_x$. Occasionally in later chapters we shall find it convenient to designate deviations from the mean by x instead of x' . If so, we shall state that the origin of x is taken at \bar{x} .

In educational work it is sometimes advisable to standardize scores, especially for the purpose of combining scores on different tests, which may show different degrees of variability. In order to avoid negative numbers, it is customary to multiply the z value by 10 and add 50 and thus obtain the standard Z score, where $Z = 10z + 50$. (A z -value below -5 seldom occurs.)

Example. (Walker). The raw scores obtained by pupils A, B, C , and D on three tests are given below. The means and standard deviations for these tests for the whole class are also given. Compute a Z score for each pupil.

Pupil	Raw Score			Composite Z
	Test I	II	III	
A	42	92	10	52
B	40	90	14	55
C	35	93	16	59
D	45	81	18	54
.	.	.	.	
.	.	.	.	
.	.	.	.	
Class \bar{x}	31.2	86.5	14.7	
Class s_x	11.5	3.6	2.4	

The z values for A on the three tests are $\frac{42 - 31.2}{11.5} = 0.939$, $\frac{92 - 86.5}{3.6} = 1.528$, and $\frac{10 - 14.7}{2.4} = -1.958$, respectively. The corresponding Z scores are 59.4, 65.3, and 30.4, the arithmetic mean of which is 51.7. This is rounded off to 52. The student should check the other values given, as an exercise.

7.8 Moments in Standard Units. When expressed in standard units, population moments are usually denoted by the Greek letter α (alpha).

Thus

$$\begin{aligned}
 \alpha_r &= \frac{1}{M} \sum_i f_i z_i^r \\
 &= \frac{1}{M} \sum_i f_i \left(\frac{x_i - \mu}{\sigma} \right)^r, \text{ by definition of } z_i, \\
 &= \frac{1}{\sigma^r} \frac{1}{M} \sum_i f_i (x_i - \mu)^r \\
 (7.23) \quad &= \mu_r / \sigma^r \text{ by (7.7)}
 \end{aligned}$$

Hence

$$(7.24) \quad \begin{cases} \alpha_1 = 0 \\ \alpha_2 = 1 \\ \alpha_3 = \mu_3 / \sigma^3 \\ \alpha_4 = \mu_4 / \sigma^4 \end{cases}$$

The α 's are all pure numbers, independent of whatever unit x may be expressed in. Thus if x is in ft, μ_3 is in ft³, but σ is also in ft, so that μ_3 and σ^3 are in the same units.

In the notation used by Karl Pearson,

$$(7.25) \quad \beta_1 = \alpha_3^2, \quad \beta_2 = \alpha_4$$

and in that of R. A. Fisher

$$(7.26) \quad \gamma_1 = \alpha_3, \quad \gamma_2 = \alpha_4 - 3$$

Corresponding to the α 's we can define standardized moments for a *sample*, namely,

$$(7.27) \quad a_r = \frac{1}{N} \sum_i f_i z_i^r, \quad z_i = (x_i - \bar{x}) / s_x$$

or

$$(7.28) \quad a_r = m_r / s_x^r = m_{r,u} / s_u^r$$

However, if we wish to estimate from a sample the moments for the population it is better to use some slightly modified moments, known as *k-statistics*.

7.9 The k -statistics. A *statistic* is a quantity calculated from the observations on a sample and used to estimate some characteristic of the parent population. These characteristics are usually *parameters*, that is, they are unknown constants which appear in the equation of the frequency curve that is assumed to represent the distribution, but which vary from one distribution of the same type to another. The population moments are parameters which occur in the equations of various commonly used frequency curves. The corresponding moments for *samples* will be approximations to the population moments, but it has been shown that, for samples which are not large, better approximations to the most important moments are pro-

vided by the k -statistics,

$$(7.29) \quad \begin{cases} k_1 = \bar{x} = m_1' \\ k_2 = N s_x^2 / (N - 1) = N m_2 / (N - 1) \\ k_3 = N^2 m_3 / (N - 1)(N - 2) \\ k_4 = N^2 [(N + 1)m_4 - 3(N - 1)m_2^2] / (N - 1)(N - 2)(N - 3) \end{cases}$$

These are estimates of μ_1' , μ_2 , μ_3 , and $\mu_4 - 3\mu_2^2$ respectively. They are said to be *unbiased* estimates because if one of them, say k_3 , were calculated for a large number of samples from a population with known moments, the mean of all the values of k_3 would be the true parameter μ_3 . If N is very large, of course, factors like $N/(N - 1)$ and $N^2/(N - 1)(N - 2)$ are practically equal to 1, and then k_2 is practically the same as m_2 , k_3 as m_3 , and k_4 as $m_4 - 3m_2^2$.

7.10 Skewness. The values of α_1 and α_2 tell us nothing about a population since $\alpha_1 = 0$ and $\alpha_2 = 1$ for all distributions. But α_3 and α_4 depend on the shape of the frequency curve, and therefore can be used to distinguish between different types. Thus

$$\alpha_3 = \frac{1}{M} \sum f(x - \mu)^3 / \sigma^3$$

is a measure of asymmetry about the mean, or *skewness*. If the values of x are distributed symmetrically about the mean, there will be for every positive value of $x - \mu$ a corresponding negative value. When these are cubed they retain their signs and cancel on addition, so that $\alpha_3 = 0$. But if the distribution has a longer tail out to the right than to the left, the positive values of $(x - \mu)^3$ usually outweigh the negative ones, so that $\alpha_3 > 0$. If the distribution has a longer tail to the left, $\alpha_3 < 0$. Since μ_3 depends on the unit of x , and since the symmetry or lack of it is not a function of the unit of measurement, we divide μ_3 by σ^3 to get the pure number α_3 . Then a curve with $\alpha_3 > 0$ is said to have positive skewness, and one with $\alpha_3 < 0$ negative skewness. These cases, along with $\alpha_3 = 0$, are illustrated in Fig. 23. For most distributions α_3 will lie between -2 and 2 .

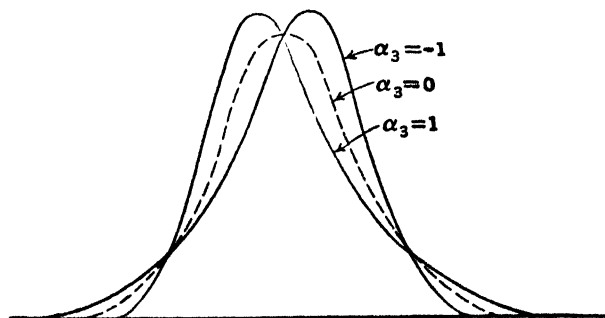


FIG. 23

With a *sample*, we can estimate the skewness of the population by using the statistic

$$(7.30) \quad g_1 = k_3/k_2^{3/2}$$

as an estimate of $\gamma_1 (= \alpha_3)$. For large N , g_1 is practically the same as $m_3/m_2^{3/2} = m_3/s^3$.

TABLE 28. THREE DISTRIBUTIONS

	A	B	C
u	f	f	f
-3	0	1	0
-2	3	1	1
-1	6	5	10
0	7	11	6
1	6	5	5
2	3	1	2
3	0	1	1
Sums	25	25	25

The data in Table 28 (simplified and adapted from actual experimental data) give frequencies corresponding to u -values for three different samples of 25. For all three the mean is zero and the standard deviation in terms of u is 1.2, but for the first two the skewness is clearly 0 (from the obvious symmetry) and for the third it is 0.74. Histograms for these three distributions are shown in Fig. 24.

7.11 Other Measures of Skewness.

For an unsymmetrical distribution the distance between the mean and mode may be used to measure the degree of asymmetry or skewness, because the mean and mode coincide in a symmetrical distribution. Since we wish any measure of skewness to be a pure number, we express this distance in units of the standard deviation, thus $(\text{mean} - \text{mode})/\sigma$.

This measure was suggested by Karl Pearson. For a certain theoretical frequency curve, known as Type III of Pearson's set of curves, the following

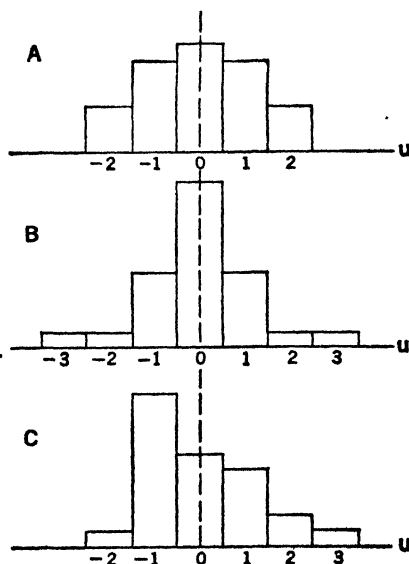


FIG. 24

relation can be proved mathematically (see Part Two, page 106):

$$(7.31) \quad (\text{mean} - \text{mode})/\sigma = \alpha_3/2$$

Because of this relation $\alpha_3/2$ is sometimes used, instead of α_3 , as the measure of skewness. Equation (7.31) can also be used for finding approximately the mode of a moderately skew distribution.

Another measure of skewness, suggested by Bowley, is based on the fact that for a positively skew distribution the third quartile is farther from the median than the first quartile, that is, $Q_3 - Q_2 > Q_2 - Q_1$. The measure adopted, which is also a pure number, is $\{(Q_3 - Q_2) - (Q_2 - Q_1)\}/(Q_3 - Q_1) = (Q_3 + Q_1 - 2Q_2)/(Q_3 - Q_1)$. This number is not as dependent as α_3 on the tails of the distribution.

7.12 Kurtosis. The fourth standardized moment, α_4 , is a measure of a property of the distribution called *kurtosis*. The name comes from a Greek word meaning "humped" and a high value of α_4 was thought to mean a sharply humped or peaked distribution and a low value of α_4 a relatively flat-topped distribution. It is now recognized that the shape of the hump has less to do with the value of α_4 than the length (and height) of the tails. Indeed, I. Kaplansky (Reference 2) has shown that kurtosis has not necessarily any-

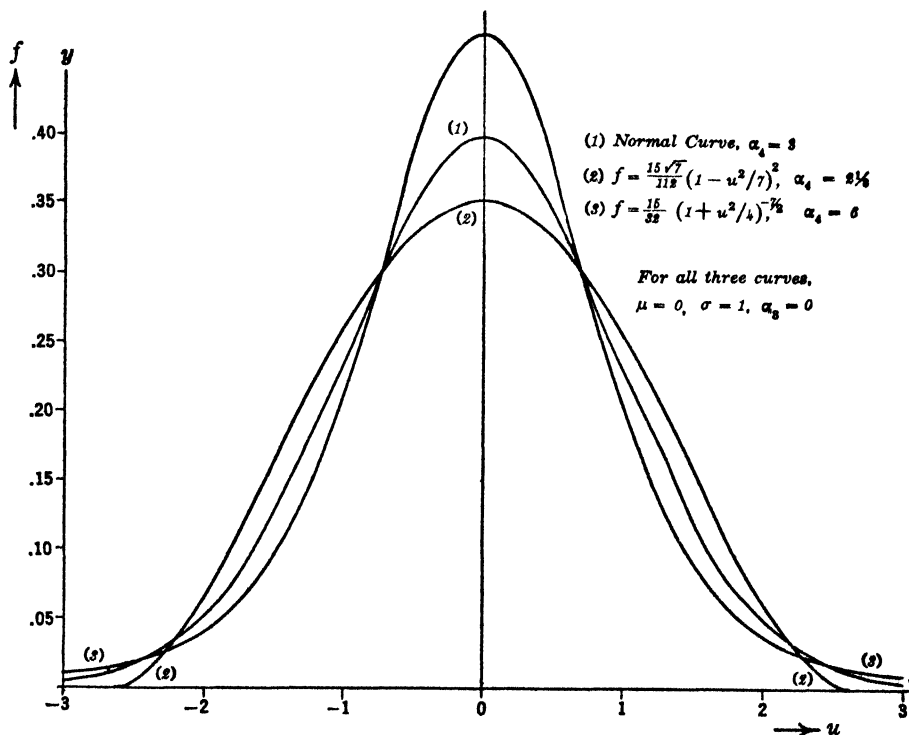


FIG. 25. KURTOSIS

thing to do with peakedness. A distribution with a perfectly flat top may have infinite kurtosis and one with an extremely sharp and high central peak may have a low value of α_4 , although these are artificially manufactured examples. For the so-called "normal curve" $\alpha_4 = 3$, so that $\gamma_2 = 0$ (see equation (7.26)), and a curve with $\alpha_4 > 3$ has positive kurtosis while one with $\alpha_4 < 3$ has negative kurtosis. Fig. 25 shows three curves with identical values of μ , σ , and α_3 , but differing in their values of α_4 .

For a *sample*, the kurtosis is measured by the statistic

$$(7.32) \quad g_2 = k_4/k_2^2$$

which is an estimate of γ_2 . For the distributions *A* and *B* of Table 28, illustrated in Fig. 24, the calculated values of g_2 are -0.85 and 1.44 respectively. For any distribution (see Reference 3)

$$(7.33) \quad \alpha_4 \geq \alpha_3^2 + 1$$

It may be shown that for a large class of theoretical frequency curves (the Pearson system, see Part Two) the mode M_0 is related to the mean μ , the standard deviation σ , the skewness γ_1 , and the kurtosis γ_2 , by the relation

$$(7.34) \quad M_0 = \mu - \frac{\sigma}{2} \frac{\gamma_1(\gamma_2 + 6)}{5\gamma_2 - 6\gamma_1^2 + 6}$$

which can be used to calculate a value for the mode.

7.13 Specimen Computation of Moments. The main characteristics of a sample distribution are summed up in the five quantities N , \bar{x} , s_x , g_1 and g_2 . In fitting any one of a considerable variety of theoretical curves to an empirical distribution, these quantities (or some of them) are used to estimate the parameters of the curve. Table 29 shows a form of work-sheet which may be used. If the work is done on a computing machine, only the totals of columns 4 to 8 need be recorded.

After obtaining \bar{u} , m_2' , m_3' , m_4' the next step is to calculate m_2 , m_3 , and m_4 by equations (7.12) to (7.14). Then the k -statistics are found by (7.29) and g_1 and g_2 by (7.30) and (7.32). All the calculations may be made in the u -variable, including the applying of Sheppard's corrections to m_2 and m_4 . The variance in terms of x is, however, c^2 times m_2 (in this example $c = 1$) and the mean in terms of x is $c\bar{u} + x_0$. We have

$$\bar{x} = 0.443 + 69.5 = 69.94 \text{ in.}$$

$$\bar{u}^2 = 0.19625, \bar{u}^3 = 0.08694, \bar{u}^4 = 0.03851$$

$$m_2 = m_2' - \bar{u}^2 = 9.901 - 0.19625 = 9.70475$$

TABLE 29. COMPUTATION OF MOMENTS OF DISTRIBUTION OF SPAN
IN INCHES AMONG ADULT MALES

x_c	f	u	uf	u^2f	u^3f	u^4f	$(u + 1)^4f$
58.5	1	-11	- 11	121	-1,331	14,641	10,000
59.5	2	-10	- 20	200	-2,000	20,000	13,122
60.5	1	- 9	- 9	81	- 729	6,561	4,096
61.5	6	- 8	- 48	384	-3,072	24,576	14,406
62.5	7	- 7	- 49	343	-2,401	16,807	9,072
63.5	22	- 6	-132	792	-4,752	28,512	13,750
64.5	55	- 5	-275	1,375	-6,875	34,375	14,080
65.5	111	- 4	-444	1,776	-7,104	28,416	8,991
66.5	146	- 3	-438	1,341	-3,942	11,826	2,336
67.5	182	- 2	-364	728	-1,456	2,912	182
68.5	229	- 1	-229	229	- 229	229	0
69.5	263	0	0	0	0	0	265
70.5	263	1	263	263	263	263	4,208
71.5	217	2	434	868	1,736	3,472	17,577
72.5	176	3	528	1,584	4,752	14,256	45,056
73.5	132	4	528	2,112	8,448	33,792	82,500
74.5	82	5	410	2,050	10,250	51,250	106,272
75.5	48	6	288	1,728	10,368	62,208	115,248
76.5	20	7	140	980	6,860	48,020	81,920
77.5	16	8	128	1,024	8,192	65,536	104,976
78.5	12	9	108	972	8,748	78,732	120,000
79.5	3	10	30	300	3,000	30,000	43,923
80.5	1	11	11	121	1,331	14,641	20,736
81.5	2	12	24	288	3,456	41,472	57,122
82.5	1	13	13	169	2,197	28,561	38,416
Sums	2,000		886	19,802	35,710	661,058	928,254
(Sums)/N			0.443 \bar{u}	9.901 m_2'	17.855 m_3'	330.529 m_4'	

Charlier's check:

$$\sum (u + 1)^4 f = \sum u^4 f + 4 \sum u^3 f + 6 \sum u^2 f + 4 \sum u f + \sum f$$

$$928,254 = 661,058 + 4(35,710) + 6(19,802) + 4(886) + 2,000$$

$$\text{Corrected } m_2 = 9.6214$$

$$s_x^2 = 9.6214, s_x = 3.102 \text{ in.}$$

$$\begin{aligned} m_3 &= m_3' - 3m_2'\bar{u} + 2\bar{u}^3 \\ &= 17.855 - 3(9.901)(0.443) + 2(0.08694) \\ &= 4.8704 \end{aligned}$$

$$\begin{aligned} m_4 &= m_4' - 4m_3'\bar{u} + 6m_2'\bar{u}^2 - 3\bar{u}^4 \\ &= 330.529 - 4(17.855)(0.443) + 6(9.901)(0.19625) \\ &\quad - 3(0.03851) \\ &= 310.432 \end{aligned}$$

$$\text{Corrected } m_4 = 310.432 - 4.852 + 0.029 = 305.609$$

$$k_2 = \frac{2000}{1999} (9.6214) = 9.6262, (k_2)^{1/2} = 3.103$$

$$k_3 = 4.8777, k_4 = 311.519 - 278.131 = 33.39$$

$$g_1 = 4.878/29.87 = 0.16, g_2 = 33.39/92.66 = 0.36$$

The estimated mean, standard deviation, skewness, and kurtosis for the population of spans are therefore 69.94 in, 3.10 in, 0.16, and 0.36, respectively.

Exercises

1. Calculate m_1' , m_3 , and m_4 for the following distributions:

(a)

(b)

x	f	x	f
0	1	-3	1
1	3	-2	3
2	5	-1	5
3	10	0	5
4	5	1	3
5	2	2	1

2. Prove the relations (7.13) and (7.14). Show that these also hold if the moments are expressed in the x -variate.

3. Prove equation (7.16).

4. Show that for any distribution expressed in standard units, the mean is 0 and the standard deviation is 1.

5. Show that the moments m_r are unaffected by a change of origin for the variable x and the moments α_r are also unaffected by a change of unit.

6. Calculate m_3 , m_4 , and m_5 for the distribution of monthly rainfall, Iowa City (Table 5, §1.8) using the scheme of Table 29.

7. Calculate the k -statistics, and g_1 and g_2 for the data of Exercise 6, according to the method of §7.13.

8. Verify the values given in §§7.10 and 7.12 for the skewness of distribution C and the kurtosis of distributions A and B , in Table 28.

9. The mean of scores for a group of students on a certain test was 63.7 with a standard deviation of 12.3. Find the Z scores for the top student, with a score of 98, and the bottom student, with a score of 21. *Ans.* 78, 15.

10. For a class of 35 students the sum of scores on a test was 2118 and the sum of squares of scores 131,327. Find the Z scores corresponding to raw scores of (a) 50 and (b) 80. *Ans.* (a) 39; (b) 71.

References

1. R. A. Fisher, "Foundation of Theoretical Statistics," *Phil. Trans. Royal Soc.*, **222A**, 1922, p. 309.

2. I. Kaplansky, "A Common Error Concerning Kurtosis," *J. Amer. Stat. Assoc.*, **40**, 1945, p. 259.

3. J. E. Wilkins, Jr., "A Note on Skewness and Kurtosis," *Ann. Math. Statistics*, **15**, 1944, pp. 333-335.

CHAPTER VIII

THE NORMAL CURVE

8.1 Frequency Curves. As we have pointed out in Chapter VII, various distributions encountered in practice can be more or less closely approximated by theoretical frequency curves. A complete discussion of such curves involves rather advanced mathematics, and fuller details will be found in Part Two (Chapter V). However, some simple ideas relating to frequency curves will be useful in our work.

The curves we shall encounter will be continuous curves, specified by an explicit mathematical equation of the form $y = f(x)$, where $f(x)$ is never negative. The domain of x is an interval of the axis, sometimes the whole axis from $-\infty$ to $+\infty$, but whether or not the curve stretches out to infinity the area between a frequency curve and the axis is always finite.

If a frequency curve is fitted to the histogram of a distribution with total frequency N , the area under the curve represents this frequency. The partial area under the curve between ordinates erected at $x = a$ and $x = b$ (Fig. 26)

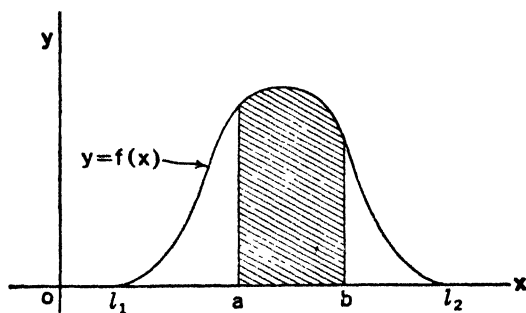


FIG. 26

represents the number of observations (the partial frequency) corresponding to a value of x between a and b . It is often convenient to consider the whole area under a frequency curve as unity, and then the area between $x = a$ and $x = b$ represents the *proportion* of observed values of x lying between a and b . This can be interpreted as meaning the *probability* that a randomly selected observation from the population represented by the curve will have a value of x between these limits; hence a frequency curve with total area unity is sometimes called a probability curve.

If the domain of x ranges from l_1 to l_2 , the total area is denoted mathe-

matically by the symbol

$$\int_{l_1}^{l_2} f(x) dx$$

which is read as "the integral of $f(x)$ from l_1 to l_2 ." It is proved in textbooks on the calculus that this integral is the limit of a sum,

$$(8.1) \quad \int_{l_1}^{l_2} f(x) dx = \lim_{\Delta x_j \rightarrow 0} \sum_j f(x_j) \Delta x_j$$

In this sum, the Δx_j are sub-intervals of the x -axis which together make up the whole interval from l_1 to l_2 , and $f(x_j)$ is the value of $f(x)$ corresponding to a point x_j in the j th interval Δx_j . Geometrically speaking, $f(x_j) \Delta x_j$ is the area of a rectangle with base Δx_j and height $f(x_j)$, and the sum is the total area of a histogram with frequencies $f(x_j)$ and class intervals Δx_j , as illustrated in Fig. 27. (The Δx_j need not be equal to each other.) The limit in

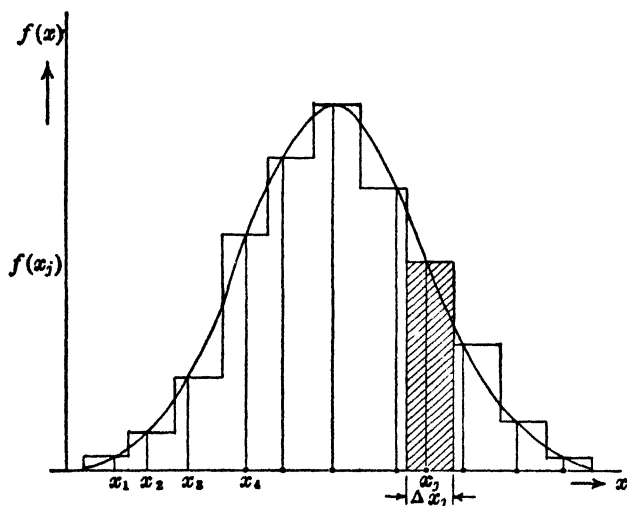


FIG. 27. ILLUSTRATING DEFINITION OF INTEGRAL

(8.1) means that all the sub-intervals are to be thought of as becoming smaller and smaller, ultimately all tending to zero. The limit may not exist, but if it does it is taken as the definition of the area under the curve $y = f(x)$ between l_1 and l_2 .

For frequency curves of the type we are considering the limit does exist, and the area under the curve is finite. The function $f(x)$ is said to be *integrable*.

The integral sign \int is merely a conventionalized S , standing for sum.

The area between the ordinates at $x = a$ and $x = b$ is similarly denoted by

$\int_a^b f(x) dx$, which, if $\int_{i_1}^{i_2} f(x) dx = 1$, represents the proportion of observations having a value of x such that $a < x < b$.

8.2 The Normal Curve. Perhaps the most important of all frequency curves is the so-called normal* curve whose equation may be written

$$(8.2) \quad y = Ke^{-h^2(x-m)^2}, \quad -\infty < x < \infty$$

where K , h^2 , and m represent numbers whose significance will be explained presently and e is the number $2.71828\cdots$ which is the base of natural logarithms. The curve is bell-shaped and is symmetrical about the line $x = m$. It was first discovered by A. De Moivre (1667–1754), a French mathematician who spent 66 years of his life in England, and it was published in 1733 in a privately printed pamphlet, now very rare. He obtained it while working on certain problems in games of chance which were proposed to him by the gamblers of his day. Because of this origin and because the data from certain coin- and dice-throwing experiments closely approach it in form, it is often called the normal probability curve. Actual statistical use of the normal curve began with the work of the famous mathematical astronomers, Laplace (1749–1827) and Gauss (1777–1855), each of whom derived it independently and presumably without knowing of De Moivre's treatment.† They found that it represented very well the errors of observation in the physical sciences. For this reason it has been called the normal curve of error, where error is used in the sense of a deviation from the true value. Since that time experience has shown that it serves quite well to describe many of the distributions which arise in the fields of biology, education, and sociology. Much of the theory of statistics is built around it.

The moments of a theoretical distribution specified by a frequency curve $y = f(x)$ can be defined by integrals.

Thus,

$$(8.3) \quad \mu_1' = \mu = \int_{i_1}^{i_2} x f(x) dx$$

$$(8.4) \quad \begin{cases} \mu_2 = \sigma^2 = \int_{i_1}^{i_2} (x - \mu)^2 f(x) dx \\ \mu_3 = \int_{i_1}^{i_2} (x - \mu)^3 f(x) dx, \text{ etc.} \end{cases}$$

where it is assumed that $\int_{i_1}^{i_2} f(x) dx = 1$.

The curve represented by (8.2) approaches the x -axis at both ends without

* The term "normal" used here should not be interpreted to mean that other types of distribution are abnormal.

† For a more extensive history, see Reference 1, page 123, and Reference 16 of §0.4.

ever quite reaching it, so that l_1 and l_2 are $-\infty$ and $+\infty$, respectively.

If the constant K is determined so that $\int_{-\infty}^{\infty} f(x) dx = 1$, and if the moments are then calculated by (8.3) and (8.4), it turns out that

$$(8.5) \quad \begin{cases} K = h/(\pi)^{1/2} \\ \mu = m \\ \sigma^2 = 1/(2h^2) \\ \mu_3 = 0 \\ \mu_4 = 3/(4h^4) \end{cases}$$

Equation (8.2) can therefore be written

$$(8.6) \quad y = \frac{1}{\sigma(2\pi)^{1/2}} e^{-(x-\mu)^2/2\sigma^2}$$

The quantities μ and σ are *parameters*. They determine the position of the curve along the x -axis and the steepness of its sides, but do not affect its general shape and character.

8.3 Standard Form. The parameters μ and σ may be removed from the equation of the curve by expressing it in terms of the standardized variable

$$(8.7) \quad z = (x - \mu)/\sigma$$

When this transformation is made, and the constant is adjusted so that the total area under the z -curve is unity, the equation of the normal curve becomes

$$(8.8) \quad \phi(z) = (2\pi)^{-1/2} e^{-z^2/2}$$

This is called the *standard form* of the equation. A variate z which is distributed in accordance with Eq. (8.8) is said to be *normally distributed with mean zero and standard deviation unity*. This is often abbreviated as $z = N(0, 1)$.

8.4 Tables of Ordinates and Areas. One of the reasons for writing the equation in standard form is that the ordinates and areas may be tabulated once and for all. These tables are given in the Appendix. We see from (8.8) that $\phi(-z) = \phi(+z)$, that is, the ordinates for negative values of z are the same as for the corresponding positive values of z , and the curve is symmetrical about the ordinate at $z = 0$. Therefore, it is not necessary to tabulate $\phi(z)$ for negative values of z .

The general shape of the normal curve may be seen from the curve (1) of Fig. 25 (§7.12). It approaches the horizontal axis asymptotically at each extremity, never quite reaching it no matter how far extended. Although an actual sample will always have a finite range it is often convenient to think of the range in the parent population as infinite and in fact this infinity leads to remarkable simplifications in more advanced mathematical statistics.

Moreover, even in representing observed distributions the infinite range causes no practical difficulty because the curve comes down to the horizontal axis very rapidly beyond $z = \pm 3$. The combined area at each extremity beyond $z = \pm 3$ is only 0.27 of 1% of the total area under the curve.

Table I of the Appendix gives also the *areas* under the standard normal curve from $z = 0$ to selected positive values of z , as far as $z = 4$. Thus, the area from $z = 0$ to $z = 1$ is 0.3413. From the symmetry of the curve the area from -1 to 0 is the same as the area from 0 to 1 . Any other areas required may be found by appropriate addition or subtraction of tabular values, remembering that the whole area under the curve from $-\infty$ to $+\infty$ is 1. For example, suppose we require the area of the "tail" of the curve below $z = -2$ (see Fig. 28) which is denoted by $\int_{-\infty}^{-2} \phi(z) dz$. The area from

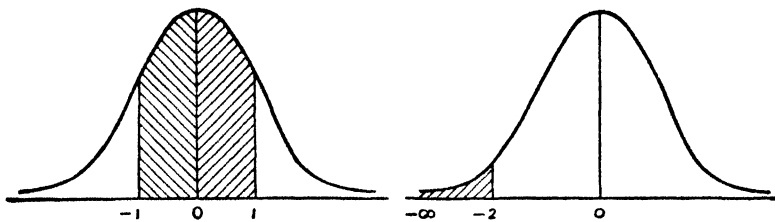


FIG. 28

$-\infty$ to -2 is 0.5 minus the area from -2 to 0 , but this latter is the same as the area from 0 to 2 , which is 0.4772. That is,

$$\int_{-\infty}^{-2} \phi(z) dz = 0.5 - 0.4772 = 0.0228$$

Note that $\int_{-\infty}^t \phi(z) dz$ is a *cumulative relative frequency*. It denotes the fraction of the total population with a value of z less than t . It is easy to see from a figure that

$$(8.9) \quad \int_{-\infty}^t \phi(z) dz = 0.5 + \int_0^t \phi(z) dz, \quad \text{or} \quad 0.5 - \int_0^{-t} \phi(z) dz$$

according as t is positive or negative. The quantity on the left of (8.9) is often denoted by $\Phi(t)$, Φ being the Greek capital phi.

For decimal values of z between the hundredths given in the table, ordinary linear interpolation will suffice.

8.5 Properties of the Normal Curve. The following properties can be established from the definition (8.8) with the help of calculus, but must for the most part be taken for granted at the level of the present course:

1. The mean, median, and mode coincide at $z = 0$. The height of the maximum ordinate is $\phi(0) = 1/(2\pi)^{1/2} = 0.3989$.

2. The curve is convex to the z -axis for $|z| > 1$ and concave to the z -axis for $|z| < 1$. The points on the curve for which $z = \pm 1$ are called *points of inflection*, and their position is important in making an accurate drawing of the curve.

3. The standard deviation is 1, and the mean absolute deviation is $(2/\pi)^{1/2} = 0.798$. The quartiles are equidistant from $z = 0$, and therefore the quartile deviation is the value of t for which

$$\int_0^t \phi(z) dz = 0.25$$

From the tables this is 0.6745. The proportion of values of z lying between -0.6745 and $+0.6745$ is 0.5000. The number 0.6745 is often, although rather ambiguously, called the "probable error" of z .

4. All the standardized moments α_r , with r odd are zero. The even moments are $\alpha_2 = 1$, $\alpha_4 = 1.3$, $\alpha_6 = 1.3 \cdot 5$, $\alpha_8 = 1.3 \cdot 5 \cdot 7$, etc.

For any other normal curve with area N , mean μ , and standard deviation σ we convert from z and $\phi(z)$ to x and y by the relations

$$(8.10) \quad \begin{cases} x = \mu + \sigma z \\ y = N\phi(z)/\sigma \end{cases}$$

The percentage distribution of area under the normal curve is indicated approximately in Fig. 29, where distances along the horizontal axis are given in units of σ . The same thing is shown in Fig. 30 with distances in units of Q (the quartile deviation).

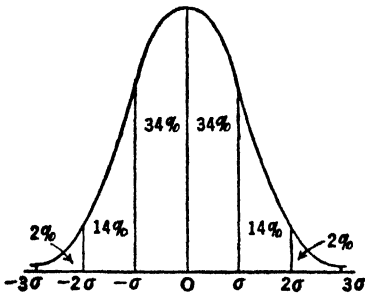


FIG. 29

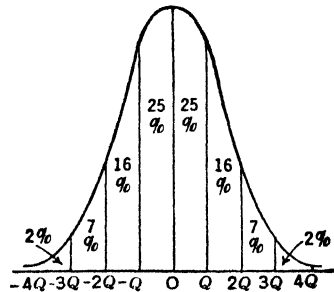


FIG. 30

8.6 Fitting a Normal Curve to a Distribution. A set of data as collected and tabulated usually relates to a sample of N individuals from a finite or infinite population. Other random samples of N from the same population would yield different frequency distributions, but if the sample is fairly large nearly all these distributions would be much alike. If the sample distribution appears to be reasonably symmetrical, bell-shaped, and tapering off

gradually at both ends, it may be worth while to try whether it can be fitted satisfactorily with a normal curve. The theoretical curve idealizes the observational data, smoothing out the irregularities due to sampling fluctuation. Furthermore, if the fit is good, the mathematical statistician can proceed to deduce various results about the behavior of samples from a normal parent population and can feel confident that his assumptions apply reasonably well to the actual population sampled.

In fitting equations (8.10) to a given distribution we *assume* that:

1. The area under the curve is equal to the area of the histogram (that is, N).
2. The mean and variance of the normal curve are unbiased estimates of the population mean and variance based on the corresponding statistics of the sample. These unbiased estimates are furnished by the k -statistics (§7.9). The estimate of μ is k_1 or \bar{x} (the sample mean), and the estimate of σ^2 is k_2 , or $Ns_x^2/(N-1)$.

The procedure of fitting a normal curve to an observed distribution will now be illustrated with the data of Table 27, §7.7, referring to weights of 1000 Glasgow schoolgirls. The calculated values of k_1 and k_2 in the original units are 47.712 and 33.344, respectively, so that we may take as parameters of the normal curve:

$$\begin{cases} N = 1000 \\ \mu = 47.712 \text{ lb} \\ \sigma = (33.344)^{1/2} = 5.7744 \text{ lb} \end{cases}$$

We now calculate standardized z values corresponding to selected values of x , by putting

$$z = \frac{x - 47.712}{5.7744} = 0.17318x - 8.2627$$

Appropriate values of x are the end values x_e and the class marks x_c . For the purpose of testing the fit, the end values of z will be required and these are given in the accompanying Table 30. The values corresponding to the class marks merely provide additional points for plotting the curve, and these have already been given in Table 27.

The ordinates of the normal curve at the given x -values are then obtained from

$$y = \frac{N}{\sigma} \phi(z) = 173.18 \phi(z)$$

using the values of $\phi(z)$ corresponding to the calculated z , as obtained from Table I in the Appendix. A smooth curve may then be drawn through the

TABLE 30. ORDINATES OF FITTED NORMAL CURVE

$$z = 0.17318 x_s - 8.2627$$

$$y = 173.18\phi(z)$$

x_s (lb)	z	$\phi(z)$	y	f/c
31.5	-2.808	0.00774	1.34	0.25
35.5	-2.115	.0426	7.38	3.50
39.5	-1.422	.1451	25.13	14.00
43.5	-0.729	.3058	52.96	43.00
47.5	-0.0367	.3987	69.05	61.25
51.5	0.656	.3217	55.71	65.75
55.5	1.349	.1606	27.81	39.00
59.5	2.042	.0496	8.59	16.75
63.5	2.734	.00951	1.65	5.75
67.5	3.427	.00112	0.19	0.75

plotted points (x, y) . Note that the mode of the curve should be at $x = 47.712$ and that the curve should be symmetrical about the ordinate through the mode.

After the curve has been drawn, the histogram for the observed data may be constructed. Since the class interval is here 4 lb, the heights of the rectangles are the frequencies divided by 4, as given in the column f/c . The values of x_s are the *ends* of the bases of the corresponding rectangles. The completed diagram is drawn in Fig. 31.

8.7 Graduation. The areas under the fitted curve and over the class intervals are called theoretical frequencies. Thus in Fig. 31 the shaded area represents the theoretical frequency corresponding to the observed frequency which is represented by the rectangle the midpoint of whose base is 41.5 lb. The determination of the theoretical frequencies is called "graduation by the normal curve." It is a process of smoothing out the data to fit the curve.

In order to enter a standard table of areas the x_s values must be changed into z values. This has already been done in Table 30. For each of these z values, we read off from Table I in the Appendix the area A under the standard normal curve from $-\infty$ up to the calculated z . For positive values of z ,

$$A = \int_0^z \phi(z) dz + 0.5$$

and for negative values of z ,

$$A = 0.5 - \int_0^{-z} \phi(z) dz$$

For example, when $z = -2.115$, we find that $\int_0^{2.115} \phi(z) dz = 0.4828$, so that $A = 0.0172$. These values of A represent *relative cumulative frequencies*.

By differencing them (see §1.11) we get the relative frequencies ΔA corresponding to the various class intervals, as set out in Table 31. The first class, however, extends from $z = -\infty$ up to $z = -2.808$, and the last class from $z = 2.734$ up to $z = \infty$. The absolute frequencies are found by multiplying ΔA by N ($= 1000$). These values of the theoretical or calculated frequencies f_c may be compared with the corresponding observed frequencies f_o , which are repeated in the last column but one of Table 31. It will be seen that there is a general similarity, although it would be hard to say whether or not the agreement is satisfactory. A method of judging this agreement will be given in a later chapter.

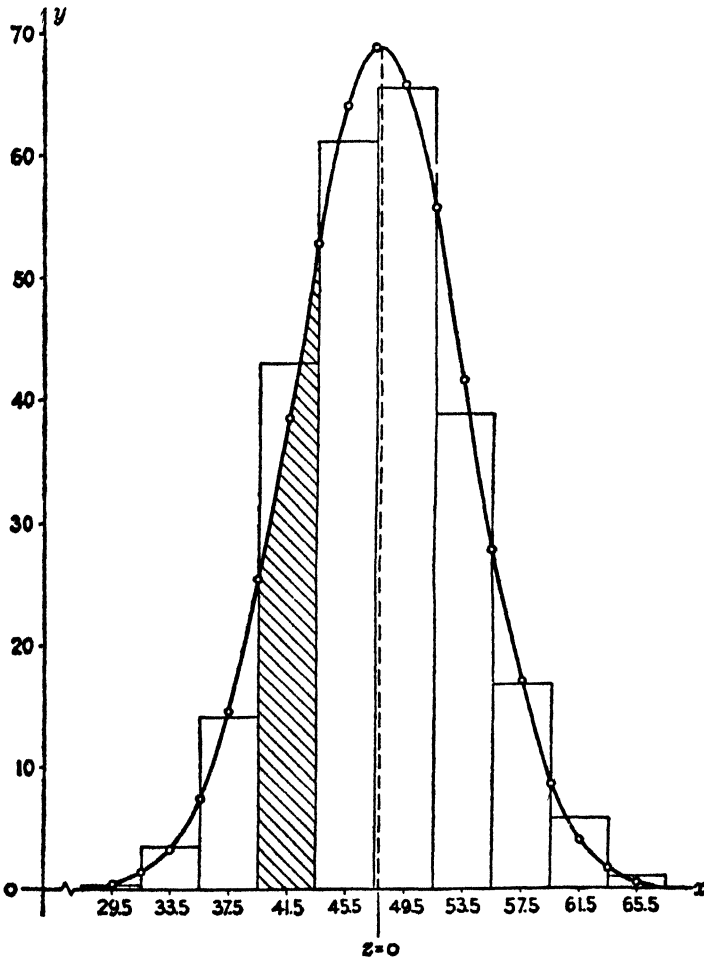


FIG. 81

TABLE 31. AREAS OF FITTED NORMAL CURVE

x_0 (lb)	z	A	ΔA	$N\Delta A = f_0$	f_0	F_0/N
	$-\infty$	0				0
31.5	-2.808	0.0025	0.0025	2.5	1	0.001
35.5	-2.115	.0172	.0147	14.7	14	.015
39.5	-1.422	.0775	.0603	60.3	56	.071
43.5	-0.729	.2330	.1555	155.5	172	.243
47.5	-0.0367	.4854	.2524	252.4	245	.488
51.5	0.656	.7441	.2587	258.7	263	.751
55.5	1.349	.9113	.1672	167.2	156	.907
59.5	2.042	.9794	.0681	68.1	67	.974
63.5	2.734	.9969	.0175	17.5	23	.997
	∞	1.0000	.0031	3.1	3	1.000
				1000.0	1000	

8.8 Justification for Using the Normal Curve. The theoretical frequency curve has the same total area, the same mean, and the same standard deviation as the observed distribution. These conditions were, in fact, imposed in the process of graduation. Furthermore, if the fitting is justifiable, the skewness and kurtosis of the frequency curve should not differ appreciably from those of the distribution itself. In the example which we have carried through in detail, we find that the estimated skewness of the population is $g_1 = 0.115$ and the estimated kurtosis is $g_2 = -0.104$. For a normal curve these values should be zero. The question arises whether these observed statistics are sufficiently near to the theoretical values to justify us in using the normal curve to graduate the data. It is proved in Part Two that, for a large value of N and for a normal parent population, the variance of g_1 among random samples of the same size is approximately $6/N$ and the variance of g_2 is approximately $24/N$. For $N = 1000$, this means that the standard deviation of g_1 among samples is about $(0.006)^{1/2} = 0.077$ and that of g_2 is $(0.024)^{1/2} = 0.155$. Now the observed g_1 differs from 0 by about $1\frac{1}{2}$ times its standard deviation and the observed g_2 differs from 0 by about $\frac{2}{3}$ of its standard deviation, and these discrepancies are quite compatible with the assumption that the true values of g_1 and g_2 are zero. The fraction of the area under a normal curve outside the range of $1\frac{1}{2}$ standard deviations from the mean is about $\frac{1}{8}$, which means that, if the values of g_1 are distributed approximately normally, there is a probability of $\frac{1}{8}$ of getting a value of g_1 at least as different from zero as 0.115. A probability of $\frac{1}{8}$ is not so small that we need reject the assumption that our sample comes from a normal distribution. (We should need a probability at least as small as $\frac{1}{10}$ and perhaps even as small as $\frac{1}{100}$ to do this.) With the kurtosis the argument is still stronger.

If we plot the values of A in Table 31 against the corresponding values of x_0 , we get a theoretical relative cumulative frequency curve, or ogive, the

characteristic shape of which is shown in Fig. 32. We can plot the observed relative cumulative frequencies $F_</math>/ N , given in the last column of Table 31, against the same values of x , and note how they will lie on the curve. The agreement can be judged more easily, however, if the scale of the graph paper$

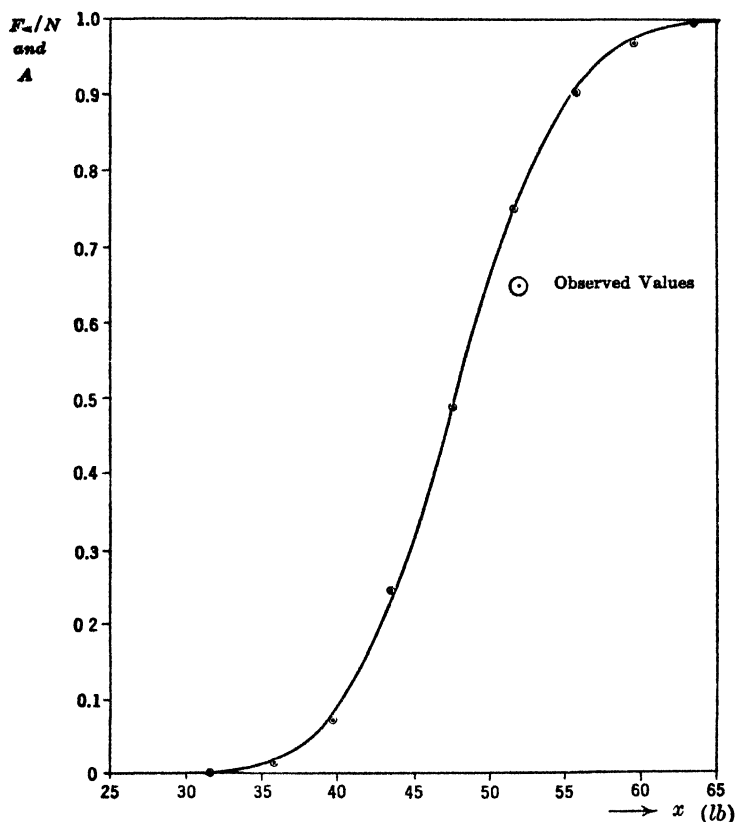


FIG. 32. OGIVE FITTED TO CUMULATIVE FREQUENCY DISTRIBUTION

is so adjusted that the ogive becomes a straight line. Imagine the vertical scale so stretched out in the neighborhood of $A = 0$ and $A = 1$ compared with the scale near $A = 0.5$ that the ogive is pulled out into a straight line. This is in effect what is done in the so-called "probability graph paper," which is illustrated in Fig. 33. This paper is readily obtainable* and is convenient for many purposes.

The plotted points in Fig. 33 are seen to lie fairly well on a straight line, which indicates that a normal curve may be expected to fit. Discrepancies near the ends of the distribution may be ignored. By drawing in the straight

* The Codex Book Company, New York.

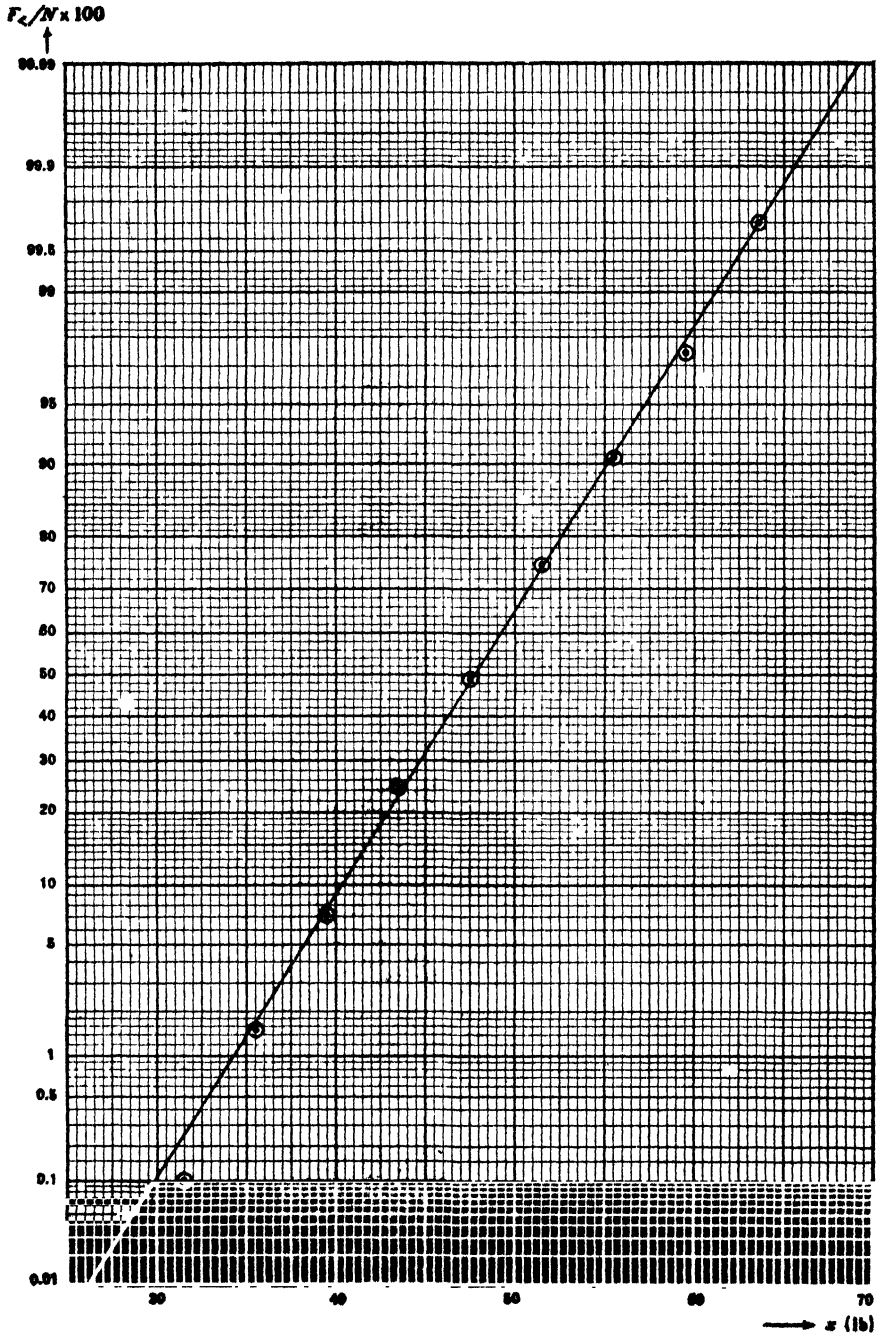


FIG. 33. OGIVE OF FIG. 32 ON PROBABILITY GRAPH PAPER

line one may make a quick rough estimate of the median, quartiles, etc., for the distribution, and estimate the theoretical frequencies lying between given values of x . For a discussion of a different type of probability paper see Reference 2.

8.9 Purpose of Graduating a Curve. Since the graduating curve is characterized by practically the same set of moments as the observed distribution, one may wonder what is the purpose of graduation. The following quotation from Prof. B. H. Camp illustrates this point (see Reference 3).

There are three main reasons why a student should be taught to graduate a curve. The first, and least important, has to do with the use of a smooth curve in place of a jagged sample. The second, and most important, is that it is necessary for the mathematical development of statistics that the mathematician should be told what assumptions he may make. These usually depend on the types of frequency curves which can be depended on to fit phenomena. . . . A third reason, intermediate in importance between the other two, is that in testing *a priori* theories in various fields, it is often necessary to test the efficacy of the frequency distributions which are results of these theories.

The second and third of Prof. Camp's reasons are not very easy to understand at the level of this book. In the theory of sampling (see Part Two) it is necessary to make assumptions about the parent population, and the mathematician naturally chooses for investigation a parent population which can be represented by a tractable mathematical function. Of all the curves which might be taken to represent reasonable frequency distributions, the normal curve has the simplest mathematical properties.

The first reason is more readily understood. Occasionally in practical problems it may be desirable to use the theoretical frequencies obtained by graduation in place of the observed data which probably contain irregularities due in part to grouping, in part to sampling fluctuations. We cite here two illustrations.

Example 1 A company which operates a chain of men's haberdashery stores planned to bring out a new line of about 100,000 lightweight sport shirts suitable for camping, hunting, etc. The question arose as to the determination of the number of each size that should be ordered from the factory. Their previous distribution of sizes had not been satisfactory because the demand for certain sizes had been different from the number manufactured. Therefore the statistical department was requested to recommend the distribution of the proposed order according to neck sizes. The solution of the problem hinged upon the availability of data giving the measurements of neck circumferences of a large sample of men. Satisfactory data were found in the "Reports of the Medical Department of the United States Army in the World War," which gave a table of the neck measurements in centimeters of 95,102 white troops at demobilization. Since these data are tabulated in class intervals which are slightly different from the ranges used in standard shirtband sizes, a slight adjustment was necessary. But essentially a normal curve was fitted to this distribution and the graduated frequencies were taken as the number of potential customers for each shirt size. The result was quite satisfactory.

Example 2. A well known and interesting illustration of the desirability of smoothing occurs in the census returns. The census takers' records show more persons alive at age 30 than at age 29, more at age 35 than at age 34, more at 40 than at 39, etc. This is probably

due to the fact that men (as well as women) do not tell their exact ages. A person who is actually 41 or 42 and known to be 40 or so, says he is 40. The recorded data show artificial bumps at every age which is a multiple of 5. Naturally the Census Bureau prefers the smoothed results to the observed. The student should not infer that the curve used to smooth these data is the normal type. The "life curve" is a continuously decreasing function. However, the same kind of quinquennial irregularity occurs in other actuarial data which do approximate the form of a normal curve. Many examples are given in Elderton, *Frequency Curves and Correlation* (Camb. Univ. Press, 4th ed., 1953).

8.10 Normalizing an Ordered Series. Suppose a large class of students are given ratings A, B, C, D or F according to an estimate of their mathematical ability based on class-work and home-work, and it is desired to give approximate scores based on the theory that mathematical ability is more or less normally distributed. This can be done by forming a relative cumulative frequency distribution and finding z -values which correspond to the dividing points between the classes. These z -values can be transformed into x -values in any convenient way by fixing two points on the scale. For example, let the frequencies in the various classes be as shown in the following table.

Rating	f	$F_{<}$	$F_{<}/N$	z	x_e
F	4	4	0.02	-2.054	29.5
D	36	40	.20	-0.842	50.0
C	92	132	.66	0.412	71.2
B	46	178	.89	1.227	85.0
A	22	200	1.00	∞	
	<hr/> 200				

The relative cumulative frequencies $F_{<}/N$ correspond to *ends* of intervals, that is, to the dividing points between the classes. The corresponding values of z are found by interpolating in Table I of the Appendix, remembering that $F_{<}/N = \int_{-\infty}^z \phi(z) dz = 0.5 + \int_0^z \phi(z) dz$, for $z > 0$

and $F_{<}/N = 0.5 - \int_0^{-z} \phi(z) dz$ for $z < 0$. Thus, for $F_{<}/N = 0.20$, the integral is 0.30. From the table we see that the integral from 0 to 0.84 is 0.29955 and from 0 to 0.85 it is 0.30234. By interpolation between these values we find that $-z = 0.8416$, so that $z = -0.8416$. The remaining z values are obtained similarly. If we now choose an x -score of 50 as the dividing line between C and D and an x -score of 85 as that between A and B, we fix the normal curve completely, since this curve has only two parameters, μ and σ , which can be determined from two independent equations. The equations are

$$(50 - \mu)/\sigma = -0.842$$

$$(85 - \mu)/\sigma = 1.227$$

from which $2.069 \sigma = 35$; or $\sigma = 16.9$, and $\mu = 64.25$. The x -score corresponding to any z is now given by

$$x = 64.25 + 16.9z,$$

and the boundary scores are as given in the table above. The F scores are below 30, the D scores between 30 and 50, and so on.

If desired, scores may similarly be obtained for the *medians* of the respective classes. Thus, the median F -score will correspond to $F_{<} = 2$, the median D -score to $F_{<} = 4 + 18 = 22$, and so on. The scores so obtained are as follows:

Rating	$F_{<}$	$F_{<}/N$	z	x
med-F	2	0.01	-2.326	24.9
med-D	22	.11	-1.227	43.5
med-C	86	.43	-0.1764	61.3
med-B	155	.775	0.7554	77.0
med-A	189	.945	1.598	91.3

"Normalized" scores obtained in this way are not to be confused with the "standardized" scores of §7.7. There is no assumption of a normal distribution with the latter type.

Exercises

1. Find by linear interpolation from Table I of the Appendix the values of $\phi(z)$ for (a) $z = 2.174$ and (b) $z = -0.625$. *Ans.* (a) 0.03755, (b) 0.32816.

2. Find the values of $\Phi(t) = \int_{-\infty}^t \phi(z) dz$ for (a) $t = 1.81$; (b) $t = -0.24$; (c) $t = 1.037$.

Hint. $\int_{-\infty}^0 \phi(z) dz = 0.5$. *Ans.* (a) 0.96485; (b) 0.40517; (c) 0.85013.

3. Find the area under the standard normal curve: (a) between $z = 1.5$ and $z = 2.5$; (b) between $z = -2$ and $z = 1.3$. *Ans.* (a) 0.06060; (b) 0.88045.

Hint. The area from 1.5 to 2.5 is the area from 0 to 2.5 minus the area from 0 to 1.5. The area from -2 to 1.3 is the area from -2 to 0 plus the area from 0 to 1.3 and the area from -2 to 0 is the same as the area from 0 to 2.

4. Show from Table I that for a normal curve with mean μ and standard deviation σ , the percentages of area outside the given ranges are as stated:

$$\text{Outside } \mu \pm \sigma = 31.74\%$$

$$\text{Outside } \mu \pm 2\sigma = 4.55\%$$

$$\text{Outside } \mu \pm 3\sigma = 0.27\%$$

Hint. Convert these ranges into z units.

5. Find the values of z , given the following values for $\phi(z)$:

(a) 0.1267, (b) 0.0335, (c) 0.0034. *Ans.* (a) ± 1.5146 ; (b) ± 2.2259 ; (c) ± 3.087 .

Hint. $\phi(1.51) = 0.12758$, $\phi(1.52) = 0.12566$. Interpolate.

6. Find t such that

$$(a) \int_0^t \phi(z) dz = 0.2746, (b) \int_{-t}^t \phi(z) dz = 0.9973, (c) \int_{-\infty}^t \phi(z) dz = 0.7500.$$

Ans. (a) 0.7541; (b) 3.00; (c) 0.6745.

7. For a normal distribution, $N = 1500$, $\mu = 75$, $\sigma = 10$. Find (a) the value of x such that $F_z = 800$; (b) how many of the N values correspond to $x < 80$.

Ans. (a) 75.83; (b) 1037.

Hints. (a) $\int_{-\infty}^z \phi(z) dz = \frac{800}{1500}$. Find z and convert to x by the relation $x = 75 + 10z$.

- (b) Find z for $x = 80$, calculate $\int_{-\infty}^z \phi(z) dz$ and multiply by 1500.

8. Find more exact values for the approximate percentages given in Fig. 30.

9. For a certain normal distribution the median is 89.0 and the first quartile 75.5. What is the standard deviation? *Ans.* 20.0.

10. Given that for a normal distribution $N = 1000$, $\mu = 20$, $\sigma = 3.5$, find (a) the value of Q_1 ; (b) the range of x corresponding to the middle 500 of the distribution; (c) the value of x corresponding to the 90th percentile.

11. If for a normal distribution $N = 300$, $\mu = 75$, $\sigma = 15$, how many values lie between $x = 60$ and $x = 70$?

12. In a college 8 grades are given, namely, A, A-, B, B-, C, C-, D, and F. On the assumptions that ability in a large class is approximately normally distributed, that the mean of the distribution lies at the boundary between B- and C, and that each grade interval corresponds to 0.8 σ , how many out of a total of 1000 should there be in each grade?

Ans. 8, 47, 157, 288, 288, 157, 47, 8.

13. It is desired to normalize scores in a class of 16 students, the order of the students having been settled by their scores on a test. Obtain the normalized scores, supposing that the lowest score is to be 30 and the highest 95.

Hint. A relative cumulative frequency table is formed as in §3.7, the students in order from lowest to highest being assigned cumulative frequencies of 0.5, 1.5, 2.5, . . . , 15.5, which are then divided by 16. Corresponding values of z are obtained from Table I.

14. (*Yule and Kendall*). A collection of human skulls is divided into three classes according to the value of a "length-breadth index" x . Skulls with $x < 75$ are classed as dolichocephalic (long-headed), those with $75 < x < 80$ as mesocephalic (medium), and those with $x > 80$ as brachycephalic (short-headed). The percentages in the three classes in this collection are 58, 38, and 4. Find approximately the mean and standard deviation of x , on the assumption that x is normally distributed. *Ans.* $\mu = 74.4$, $\sigma = 3.23$.

15. Values of the skewness (g_1) and kurtosis (g_2) were worked out in §7.13 for the population of spans of adult males. Are these values reasonably near to zero, according to the criterion discussed in §8.8? Write down the equation of the best-fitting normal curve for this population.

16. Graduate by means of a normal curve the distribution of lengths of telephone calls, Table 25, p. 87. (Take $\mu = 477.3$ sec, $\sigma = 1.5.7$ sec.)

17. A distribution of weekly wages for 906 miners at a certain date showed the following results: $\bar{x} = \$36.13$, $s_x = \$8.87$, $a_3 = 0.607$, $a_4 = 3.02$. Assuming that the true distribution is approximately normal with the mean and standard deviation estimated from the sample, calculate the proportion of miners who received weekly wages (a) in excess of \$65, and (b) less than \$25.

18. An urban electric railway company operating a large subway uses thousands of electric light lamps in its underground stations. On January 1, 1954, the company put into

service 5000 new lamps. Assume that the distribution of length of life for these lamps is normal, with a mean of 50 days and a standard deviation of 19 days. If January 1 is counted as a full day, how many lamps out of the 5000 new ones would need to be replaced by midnight January 31, 1954? How many by March 10, 1954?

19. (Camp). The standard deviation of a set of 100,000 high school grades was 11%, and the mean grade was 78%. Assuming that the distribution was normal, find (a) how many grades were above 90%, (b) how many were below 70%, (c) what was the 99th percentile, (d) what was the semi-interquartile range.

20. In a certain normal distribution we have $N = 1000$, $\mu = 50$, $\sigma = 10$. For this distribution (a) convert the values $x = 20, 30, 40, 50, 60, 70, 80$ into the corresponding z 's; (b) find the corresponding values of $\phi(z)$; (c) convert these into y values; (d) plot the points (x, y) and sketch a smooth curve joining them; (e) calculate the partial frequencies between 20 and 30, between 40 and 50, and between 42 and 74; (f) find the values of x for which $F_x = 250, 600, 750$, respectively.

21. Suppose a variate v is normally distributed with mean 0 and variance 25.

(a) Give the equation of the frequency curve for a population of size N .

(b) If there are 793 values between $v = -5$ and $v = 0$, find N .

(c) Find what percentage of values correspond to $v > 10$.

(d) Find the value of v for which $F_v/N = 0.75$.

22. 100 individuals are graded in 5 classes, ranging from A (the highest) to E (the lowest), the frequency distribution being as follows:

Grade	A	B	C	D	E
f	5	21	39	28	7

If the grades are normalized and scores are given so that 90 is the median score for class A and 10 the median score for class E, find what scores correspond to the points of division between the classes.

References

1. H. M. Walker, "Bicentenary of the Normal Curve," *J. Amer. Stat. Assoc.*, **29**, 1934, pp. 72-75.

2. F. C. Martin and D. H. Leavens, "A New Grid for Fitting a Normal Probability Curve to a Given Frequency Distribution," *J. Amer. Stat. Assoc.*, **26**, 1931, pp. 178-183.

3. H. C. Carver, "The Concept and Utility of Frequency Distributions," *J. Amer. Stat. Assoc.*, **26**, 1931 (Supplement), pp. 33-36. Discussion on above, by B. H. Camp, p. 36.

CHAPTER IX

PROBABILITY

9.1 Meaning of Probability. The notion of probability has been introduced several times in previous chapters, but without any very precise definition. Indeed, it is extremely difficult to give a precise definition at an elementary level. The notion is so important, however, and will recur so frequently in later chapters, that we must attempt some further explanation, and indicate how simple calculations connected with probabilities can be made.

From the statistical point of view, the notion of probability is a rather natural extension of the notion of *relative frequency*. In 100 throws with a die, the 6-spot turns up, say, 15 times. We try again and this time we throw 17 sixes. If we keep on for thirty or forty sets of 100 throws each, we shall obtain a set of relative frequencies which will cluster around the values 0.16 and 0.17, and it is likely that the mean of these relative frequencies will be quite close to $1/6$. It is a matter of common experience that when we have a well-defined process which can be observed over and over again (like throwing a die) the relative frequencies of some definite characteristic associated with the process (like the number 6) always show this tendency to cluster around a fixed value, and they do so more closely the greater the number of observations. This fixed value is called the *probability* of the characteristic, and it is obviously a number between 0 and 1 inclusive. This number can be thought of as the ideal or theoretical frequency of an *event*, the event being the appearance of the characteristic in question.

Although the theory of probability started from discussions of games of chance, and although these games still furnish useful illustrations, probability has wide applications in many different fields. The "event" may, for example, be the appearance of a defect in an article turned out in large numbers by a certain factory machine, or it may be the occurrence of color-blindness in an adult male, or the possession of a taxable income of more than \$6000. There is in each case a population, finite or infinite, relative to which the probability is defined. Thus, samples of 100 machined articles may be examined daily for defects, and, as long as the machine is properly adjusted, the relative frequencies will lie near a number which is the probability that this machine will turn out a defective article. Of course, something may go wrong with the adjustments, and the machine may suddenly start turning out large numbers of defectives, but then we say that the process is "out of control." The basic conditions have changed, and we are no longer dealing with the same population. The population of adult males in, say, Canada, is not

infinite, but it is a large number, and there is a probability that if one were selected at random and examined he would turn out to be color-blind. This probability can be assessed from the known relative frequency of color-blindness among groups which have been tested, such as recruits for the Air Force. The relation of probabilities to relative frequencies is something like the relation of the ideal points and lines of geometry to the actual chalk or pencil marks made on the blackboard or on paper. These marks are a crude approximation to the ideal indefinitely small points and indefinitely thin lines about which we reason. In mechanics, again, we discuss the properties of "particles" and "rigid bodies," which are nonexistent abstractions, but the practical value of theoretical mechanics lies in the fact that many observable objects in the real world behave very much like these abstractions. How far the calculus of probabilities applies to the real world can be determined only by observation and experiment.

9.2 Combination of Relative Frequencies. Probabilities are assumed to obey certain laws of combination suggested by the corresponding laws for relative frequencies. To illustrate these laws, let us consider a simple two-way frequency distribution, known as a two-by-two (2×2) table, such as Table 32, which gives data on the incidence of a certain disease among a group of

TABLE 32. 2×2 TABLE OF EFFECT OF INOCULATION

	<i>Inoculated</i>	<i>Not-inoculated</i>	<i>Total</i>
Attacked	2	10	12
Not-attacked	5	3	8
Total	7	13	20

20 people, some of whom had been inoculated with a drug and others not. The frequencies in the vertical margin, 12 and 8, form a *marginal frequency distribution* of attack; those in the horizontal margin, 7 and 13, form a marginal frequency distribution of inoculation. In both cases there are two classes in the distribution, and the marginal frequencies are given by adding the individual frequencies, either along the rows or up the columns of the table.

The first column by itself gives a distribution of incidence of attack among the inoculated; this is called a *conditional distribution*. Similarly the second column gives a conditional distribution for the not-inoculated, and the two rows give conditional distributions of inoculation among the attacked and among the not-attacked.

The conditional relative frequency of attack among the inoculated is $2/7$, and the relative frequency of inoculation in the sample is $7/20$. The relative

frequency of individuals in the sample who are both inoculated and attacked is $2/20$, which is the product of $2/7$ and $7/20$. Again, if we want the total frequency of individuals who have *either* been inoculated *or* not attacked by the disease we can compute it by adding the marginal frequencies for inoculation and non-attack and subtracting the frequency of individuals coming under both categories, namely, $7 + 8 - 5 = 10$. This is the same as the total sample frequency less the number who are *both* attacked *and* not inoculated, that is, $20 - 10$.

These results can be generalized as follows:

Let A and B be two events and \bar{A} and \bar{B} the *complementary* events. That is, \bar{A} (read "A-tilde") denotes the event " A does not occur." For example, if A means attacked, \bar{A} means not-attacked, and every individual in the sample may be placed in one of these two classes. Similarly, every individual is either a B or a \bar{B} (namely, a B or a not- B). The generalized 2×2 frequency table is shown as Table 33, where f_{11} means the frequency of simultaneous

TABLE 33. GENERALIZED 2×2 TABLE

	A	\bar{A}	Total
B	f_{11}	f_{12}	r_1
\bar{B}	f_{21}	f_{22}	r_2
Total	c_1	c_2	N

occurrence of both A and B , etc. The marginal frequencies for A and \bar{A} are the column totals c_1 and c_2 ; the marginal frequencies for B and \bar{B} are the row totals r_1 and r_2 ; and N is the overall total frequency. Clearly, $N = c_1 + c_2 = r_1 + r_2$. We now use a notation borrowed from the subject of Symbolic Logic, in order to specify certain compound events. By AB we shall mean the simultaneous occurrence of *both* events A and B . By $A + B$ we mean the occurrence of *either* A *or* B *or both*, which may be expressed as " A and/or B ." From this definition it follows that

$$(9.1) \quad A + B = A\bar{B} + \bar{A}B + AB$$

(The reason for using mathematical notation in this sense will appear shortly.) Writing the corresponding relative frequencies as $f\{A\}$, $f\{AB\}$, etc., we see from Table 33 that

$$(9.2) \quad \begin{cases} f\{A\} = c_1/N, & f\{\bar{A}\} = c_2/N \\ f\{AB\} = f_{11}/N \\ f\{A + B\} = (f_{11} + f_{12} + f_{21})/N \end{cases}$$

It follows immediately that

$$(9.3) \quad f\{A\} + f\{\bar{A}\} = 1$$

Since $f_{11} + f_{21} + f_{12} = N - f_{22}$,

$$(9.4) \quad f\{A + B\} = 1 - f\{\bar{A}\bar{B}\}$$

Also, $f_{11} + f_{21} = r_1$, and $f_{11} + f_{12} = c_1$, so that $f_{11} + f_{21} + f_{12} = r_1 + c_1 - f_{11}$. Hence

$$(9.5) \quad f\{A + B\} = f\{A\} + f\{B\} - f\{AB\}$$

The observations in the first column of Table 33 form a conditional frequency distribution for B among those individuals, c_1 in number, for which A is known to occur. The event " B occurs when it is given that A also occurs" is denoted by $B|A$, (read " B , given A ") and from Table 33 we have

$$(9.6) \quad f\{B|A\} = f_{11}/c_1$$

Since $f\{AB\} = f_{11}/N$ and $f\{A\} = c_1/N$, we see that

$$(9.7) \quad f\{AB\} = f\{A\} f\{B|A\}$$

and similarly

$$(9.8) \quad f\{AB\} = f\{B\} f\{A|B\}$$

Relations (9.5), (9.7), and (9.8) form the basis of the axioms of probability given in the next section.

9.3 Rules for Combining Probabilities. Since we have defined a probability as a theoretical or idealized relative frequency, it is natural to assume that probabilities will obey the same rules of combination as relative frequencies. We therefore take it as axiomatic that

(1) The probability of an event A , denoted by $P\{A\}$, is a number between 0 and 1 inclusive, an impossible event having probability 0 and one that is certain to occur having probability 1.

(2) The probability that *at least one* of the two events A and B occurs is given by

$$(9.9) \quad P\{A + B\} = P\{A\} + P\{B\} - P\{AB\}$$

(3) The probability of the simultaneous occurrence of both A and B is given by

$$(9.10) \quad P\{AB\} = P\{A\} P\{B|A\} = P\{B\} P\{A|B\}$$

If the two events A and B are *mutually exclusive* (meaning that it is impossible for both to occur together), $P\{AB\} = 0$, and (9.9) becomes

$$(9.11) \quad P\{A + B\} = P\{A\} + P\{B\}$$

This is called the *addition theorem* for probabilities and is the reason for the use of the $+$ sign to denote "and/or" which, when A and B are mutually exclusive, reduces to the strict alternative "either-or."

Since the events A and \bar{A} are mutually exclusive, and one of these must occur, we have

$$(9.12) \quad P\{A\} + P\{\bar{A}\} = 1$$

The events A and B are said to be *independent* if the probability of occurrence of A is the same whether B occurs or not, that is, if

$$P\{A|B\} = P\{A|\bar{B}\}$$

Now

$$\begin{aligned} P\{A\} &= P\{AB\} + P\{A\bar{B}\} \\ &= P\{B\} P\{A|B\} + P\{\bar{B}\} P\{A|\bar{B}\} \end{aligned}$$

by (9.10). If $P\{A|B\} = P\{A|\bar{B}\}$, this becomes

$$\begin{aligned} P\{A\} &= P\{A|B\} (P\{B\} + P\{\bar{B}\}) \\ &= P\{A|B\} \end{aligned}$$

and therefore (9.10) can be written, for independent events, .

$$(9.13) \quad P\{AB\} = P\{A\} P\{B\}$$

This is called the *multiplication theorem* for probabilities, and is the reason for using AB to mean " A and B ."

These theorems may be generalized to any number of events. Thus, if the events A_1, A_2, \dots, A_m are mutually exclusive and if one of them must occur,

$$\begin{aligned} (9.14) \quad P\{A_1\} + P\{A_2\} + \dots + P\{A_m\} &= \\ P\{A_1 + A_2 + \dots + A_m\} &= 1 \end{aligned}$$

Let us now suppose that the probability is the same for each of these m events, say P . Then $mP = 1$, so that $P = 1/m$.

If we designate the first k of the m events as "favorable," the probability of a favorable event is

$$\begin{aligned} (9.15) \quad P\{A_1 + \dots + A_k\} &= P\{A_1\} + P\{A_2\} + \dots + P\{A_k\} \\ &= k/m \end{aligned}$$

This result was used as the *definition* of probability by Laplace and the early writers on the subject. An example is the throwing of a die, discussed in §9.1. Here there are six mutually exclusive possibilities and one of them, the turning-up of the 6-spot, is regarded as favorable, so that $k = 1$ and $m = 6$. If we consider that any one of the six faces of the die is precisely as likely to turn up as any other, the probability of throwing 6 is $\frac{1}{6}$. This definition of probability is convenient in some situations, particularly in connection

with games of chance, but its applicability is very limited, and in most statistical problems the frequency definition makes much more sense. For a further discussion see Chapter I of Part Two.

9.4 Permutations. For a considerable number of simple calculations in probability, based on equation (9.15), it is necessary to make use of the elementary algebra of permutations and combinations, and we shall now develop some of this theory.

A *permutation* of a finite number of distinguishable objects is any arrangement of these in a definite order. For example, the objects may be thought of as numbered wooden blocks. With two blocks there are two possible arrangements 1 2 and 2 1. With three blocks there are six arrangements, namely 1 2 3, 1 3 2, 2 1 3, 2 3 1, 3 1 2, and 3 2 1. The student can easily verify, by writing them out, that there are 24 different arrangements of the numbers 1, 2, 3, 4. We can get a general formula with n blocks, by considering that the first place in the ordered arrangement can be filled in n ways, since any one of the n blocks can be chosen. The next place can be filled in $n - 1$ ways, since only $n - 1$ blocks are left to be picked. Similarly the third place can be filled in $n - 2$ ways, and so on, until the last place can be filled in only one way, with the last block left. The total number of ways is, therefore,

$$n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$$

and this number is denoted by $n!$ (read “ n factorial”). For $n = 4$, $n! = 24$, for $n = 5$, $n! = 120$, and so on.

This principle of filling places one at a time is often useful in working problems.

Example 1. There are 6 seats available in a car. In how many ways can 6 persons be seated for a journey, if only 3 of them can drive?

Here the first place may be taken as the driver's seat, which can be filled in 3 ways. For the next seat there are 5 persons available, for the next 4, and so on. The total number of ways is therefore $3 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 360$.

Theorem 1. *The number of ways of selecting r objects out of n distinguishable objects, and arranging them in order, is*

$$(9.16) \quad P(n, r) = n(n-1)\cdots(n-r+1) = n!/(n-r)!$$

This is an application of the same principle. The first place can be filled in n ways, the second in $n - 1$, and the r th in $(n - r + 1)$ ways, so that $P(n, r) = n(n-1)\cdots(n-r+1)$. If we multiply $P(n, r)$ by $(n-r)!$ we get $n(n-1)\cdots(n-r+1)(n-r)(n-r-1)\cdots 1$, which is $n!$, thus verifying the second form given in (9.16). $P(n, r)$ is usually called the number of permutations of n things, r at a time. Observe that $P(n, r)$ denotes an integer.

Theorem 2. *The number of ways of filling r places with objects selected out of n distinguishable objects, when the same object can be used as often as desired, is n^r .*

This follows at once from the fact that each place can be filled in n ways, regardless of the way the earlier places have been filled.

Example 2. The number of 5-digit license plate numbers that can be formed from the ten digits 0, 1, \dots , 9 is $10^5 = 100,000$. If two letters of the alphabet followed by three digits are used, the number is $26^2 \cdot 10^3 = 676,000$.

Theorem 3. *If p out of n objects are indistinguishable from each other, the number of permutations is $n!/p!$.*

Proof: The total number of permutations is $n!$, but for every arrangement of the $n - p$ different objects there are $p!$ permutations which are identical, because they differ among themselves only by rearrangement of the p indistinguishable objects. The number of different permutations is therefore $n!/p!$.

Corollary. *The number of permutations of n objects, of which n_1 are alike of one kind, n_2 alike of another kind, and so on, is $n!/(n_1!n_2!\cdots n_k!)$, where $n_1 + n_2 + \cdots + n_k = n$.*

Example 3. The number of arrangements of the letters of the word "independent," taken all together, is $11!/(3!3!2!) = 554,400$, since there are 11 letters including 3 e's, 3 n's and 2 d's.

9.5 Combinations. If we want to know in how many ways we can pick out a number of objects from a collection, *not caring in what order they are arranged*, we have a problem in *combinations*. Such problems are much more important in probability theory than those in permutations, because we are seldom interested in arrangements as such. The number of ways of picking out r objects from n distinguishable objects, called *the number of combinations of n things r at a time*, will be denoted by $C(n, r)$ or by $\binom{n}{r}$ which may be read " n above r ." The latter notation is now common, and other notations such as nC_r are also used.

Theorem 4.

$$(9.17) \quad C(n, r) = \frac{n!}{r!(n - r)!}$$

By permuting each combination of r things among themselves we shall obtain all possible permutations of n things, r at a time. Each combination gives rise to $r!$ permutations, so that $r!C(n, r) = P(n, r) = n!/(n - r)!$, whence the theorem follows.

Corollary.

$$(9.18) \quad C(n, r) = C(n, n - r)$$

This follows from (9.17) by writing $n - r$ instead of r . It is also obvious, since, if we pick r things out of n , we pick at the same time the $n - r$ things which are left behind.

Theorem 5.

$$(9.19) \quad C(n, n) = 1$$

There is clearly only one way of picking all the n objects, so that $C(n, n)$ must be 1. If equation (9.17) is still to apply when $r = n$, we must interpret the symbol $0!$ as 1, and this is the convention that is followed. Similarly we must interpret $C(n, r)$, for $r > n$, as being zero.

Example 4. In how many ways can a committee of 3 be chosen from 5 married couples, if a husband and wife cannot both sit on the committee?

We can pick three couples out of the five in $C(5, 3) = 5!/(3!2!) = 10$ ways. A member can be chosen from each couple in $C(2, 1) = 2$ ways. The total number of ways is therefore $10 \cdot 2 \cdot 2 \cdot 2 = 80$.

Theorem 6. If n objects consist of n_1 all of one kind, n_2 all of another kind, and so on, up* to n_k of the k th kind, the total number of selections that can be made of 1, 2, 3 up to n objects is

$$(9.20) \quad (n_1 + 1)(n_2 + 1) \cdots (n_k + 1) - 1$$

We may take either none or 1 or 2 or up to n_1 of the first kind, giving $n_1 + 1$ possibilities. Similarly we may take none or 1 or 2 or up to n_2 of the second kind, giving $n_2 + 1$ possibilities, and so on. But we exclude the case when we select none of any kind.

Corollary. The total number of selections from n objects all different is $2^n - 1$.

This is found from (9.20) by putting $n_1 = n_2 = \cdots = n_k = 1$, and noting that $k = n$.

Example 5. A traveler has in his pocket a nickel, a dime, a quarter, and a half-dollar. In how many ways can he give the porter a tip? *Ans.* $2^4 - 1 = 15$.

Theorem 7. The number of ways of putting m indistinguishable objects into n numbered compartments, if any number (including 0) may go into any compartment, is

$$C(n + m - 1, m) = (n + m - 1)! / [m!(n - 1)!]$$

Proof: We think of the m objects as arranged in a line, thus:

$$0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \cdots 0$$

Now we imagine $n - 1$ vertical lines placed anywhere between these objects, thus:

$$0|0 \ 0||0 \ 0 \ 0|0|\cdots|0$$

These partitions will separate n compartments, which will contain none or 1 or 2 or up to m of the objects, and these compartments may be supposed

* "Up to n " includes n .

numbered from left to right. The number of possible arrangements will be the number of permutations of $m + n - 1$ objects, of which m are alike of one kind and $n - 1$ alike of another kind (namely, the partitions). By the corollary to Theorem 3, this is $(m + n - 1)! / [m!(n - 1)!]$.

Example 6. There are 7 horses in a race. If a man has 5 dollars to bet with, and can bet only in multiples of a dollar, in how many ways can he bet on one or more horses to win?

Here we have 5 objects (dollars) to put into 7 numbered compartments (ticket windows). The number of ways is $C(11, 5) = 462$.

9.6 Some Problems in Probability. We now give a few examples of problems which can be solved by enumerating the possible cases and the favorable cases, and using equation (9.15).

Example 7. What is the probability of holding 4 aces in a hand at bridge?

The number of possible hands of 13 cards is $C(52, 13)$. The number of hands containing 4 aces is equal to the number of ways that the remaining 9 cards can be picked out of the 48 cards in the deck which are not aces. This number is $C(48, 9)$. The required probability is, therefore,

$$\frac{C(48, 9)}{C(52, 13)} = \frac{48! 39! 13!}{9! 39! 52!} = \frac{13 \cdot 12 \cdot 11 \cdot 10}{52 \cdot 51 \cdot 50 \cdot 49} = 11/4165$$

or about 1 in 379.

The fundamental assumption here is that every *completely specified* hand is as likely as any other one, dealt from a well-shuffled deck. The hand consisting of all 13 spades is just as likely as the hand consisting of clubs, K, 10, 3, diamonds A, J, 5, 4, hearts Q, 9, 8, and spades 10, 7, 4. There are, however, a very large number of different specified hands, about 635 billion in fact, and only 4 of these are hands consisting of a complete suit, so that the probability of such a hand is extremely small. In actual play the shuffling is seldom very thorough and it may be doubted whether the fundamental assumption is justified.

Example 8. Find the probability of a hand of 13 cards containing 3 clubs, 4 diamonds, 3 hearts, and 3 spades.

The number of ways of picking 3 clubs out of the 13 clubs in the deck is $C(13, 3)$. Each of these ways may be associated with any of the $C(13, 4)$ ways of picking the diamonds, the $C(13, 3)$ ways of picking the hearts, and the $C(13, 3)$ ways of picking the spades. The total number of favorable ways is, therefore, the product of these numbers, and since the total number of ways is $C(52, 13)$, the required probability is

$$[C(13, 3)]^3 C(13, 4) / C(52, 13) = \frac{(13!)^3 13! 13! 39!}{(10!)^3 (3!)^3 9! 4! 52!} = 0.0263$$

9.7 Simple and Compound Events. A simple event is one which can be represented by some value or set of values of a single variate x . In some cases x is discrete, that is, it can take only isolated values such as 0, 1, 2, 3, In other cases it may range over a finite or infinite interval of the x -axis. The throwing of a die and the observation of the number of spots on the upper face constitute a simple event, where x can take only the values 1, 2, 3, ..., 6. Another simple "event" would be an adult male having a height of over 6 ft, and x would then be any number greater than 6. The event in Example 7, that of having a certain number of aces in a hand of bridge, can be represented by $x = 0, 1, 2, 3$, or 4.

For a *compound* event we require two or more variates. Thus, if two dice are thrown together, any number of spots from 1 to 6 on the first die may be associated with any number from 1 to 6 in the second die. There are therefore 36 possible combinations, which can be represented by points (x, y) in a plane, where x and y both take the values 1, 2, \dots , 6. These points form a square arrangement of isolated dots. By writing out all the possible combina-

(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1	1	2	1	3	1	4	1	5	1	6	1		
1	2	2	2	3	2	4	2	5	2	6	2		
1	3	2	3	3	3	4	3	5	3	6	3		
1	4	2	4	3	4	4	4	5	4	6	4		
1	5	2	5	3	5	4	5	5	5	6	5		
1	6	2	6	3	6	4	6	5	6	6	6		

tions in a table, we can see at a glance how many correspond to a given total. Thus there is only one combination, 6 6, which will give a total of 12 and the probability of this total is $\frac{1}{36}$, if we assume that the dice are uniform and well-balanced so that all faces are equally likely to appear. A total of 7 can, however, be made up in 6 different ways: 1 6, 2 5, 3 4, 4 3, 5 2, and 6 1, so that the probability of a total of 7 is $\frac{1}{6}$. One can check easily that the probabilities of the various possible totals are as given in the following table:

Total	2	3	4	5	6	7	8	9	10	11	12
Prob.	1	2	3	4	5	6	5	4	3	2	1

($\times 36$)

The probabilities are multiplied by 36 to avoid fractions. The sum of the probabilities is 1, as it should be.

Example 9. What is the probability of throwing either 7 or 11 with two dice?

These events are mutually exclusive and the probability is the sum of the individual probabilities. The probability of 7 is $\frac{1}{6}$ and that of 11 (from the above table) is $\frac{1}{18}$. The sum is $\frac{5}{18}$.

Example 10. Show that the probability of throwing 6 at least once in 4 throws of a die is a little more than 0.5, but that the probability of throwing double-6 at least once in 24 throws with two dice is a little less than 0.5.

The probability that an event happens at least once is the complement of the probability that it does not happen at all. If we calculate the latter, we have only to subtract it from 1 to get the former. The probability of *not* getting 6 in a single throw is $\frac{5}{6}$. Since the successive throws of the die are supposed to be independent, the probability of *not*-6 in four throws is $(\frac{5}{6})^4$. Therefore the probability of at least one six is

$$1 - (\frac{5}{6})^4 = 671/1296 = 0.516$$

Similarly, the probability of *not* getting double-6 in a single throw with two dice is $\frac{35}{36}$. The probability of *not* getting it in 24 throws is $(\frac{35}{36})^{24}$ and the probability of at least one double-6 is

$$1 - (\frac{35}{36})^{24} = 0.491$$

This problem was one of the earliest to be solved in the history of the theory of probability. The Chevalier de Méré, a gambler at the French court in the middle of the seventeenth century, had noticed that, while it paid to bet on the first event, it did not pay to bet on the second. This seemed to him unreasonable, and he consulted the mathematician Pascal (1623-1662) who worked out the probabilities.

9.8 Continuous Probability. The problems so far considered have been examples of discrete variates, when the number of possible values was finite. There are other problems, however, where x is a *continuous* variate, so that no enumeration of favorable and total cases is possible. If we consider first a simple event, there will be, in general, a probability that x will lie in a specified interval out of its whole domain, and this probability will be a function of x . There is, for example, a probability that an adult Canadian male, selected at random, will have a height less than or equal to 6 ft. The function $F(x)$ which expresses the probability that a certain variate has a value equal to or less than x is called the *distribution function* of the variate. If the variate follows the normal law, for example, the graph of the distribution function is an ogive like that in Fig 32 of §8.8. For distributions expressed by smooth continuous curves, there will also be a probability that x lies in an infinitesimal interval of the domain, denoted in the language of calculus by dx . Since this probability will be proportional to the size of dx , we denote it by $f(x) dx$, and we call $f(x)$ the *probability density*, or the *probability function*, of the distribution. If the whole domain of x stretches from l_1 to l_2

$$\int_{l_1}^{l_2} f(x) dx = 1$$

and if we regard as a "favorable" event one which corresponds to a value of x between certain limits k_1 and k_2 , the probability of such an event is given by

$$\int_{k_1}^{k_2} f(x) dx$$

In many problems it is reasonable to consider that every value of x within its domain is *equally likely*. If so, $f(x)$ is constant, and we can easily evaluate the probability. If $f(x) = C$,

$$\int_{l_1}^{l_2} f(x) dx = C(l_2 - l_1)$$

so that $C = 1/(l_2 - l_1)$.

Then

$$\int_{k_1}^{k_2} C dx = C(k_2 - k_1) = (k_2 - k_1)/(l_2 - l_1)$$

Thus if a point is selected at random on a line 6 inches long, the probability that it lies within an inch either way of the middle point is $\frac{2}{6} = \frac{1}{3}$, since the favorable interval is 2 inches long and the whole domain 6 inches long.

This principle may be extended to compound events. If, for example, a favorable event corresponds to the position of a point (x, y) within a certain region R of the x - y plane, and if the whole domain of x and y is a region D , then, on the assumption that all positions of the point are equally likely, the probability of the favorable event is R/D .

Example 11. A horizontal flat plate 8 in. square is ruled with a grid of fine lines, 2 in. apart, and has a vertical rim all around the edge. If a penny (diameter $\frac{3}{4}$ in.) is tossed on to the plate, what is the probability that it rests without crossing a line?

Because of the rim the center of the penny cannot lie within $\frac{3}{8}$ in. of the edge of the plate. Hence, the effective domain D is a square of side $7\frac{1}{4}$ in. If the penny is not to cross a line of the grid, its center cannot lie within $\frac{3}{8}$ in. of any such line. The favorable region, therefore, consists of 16 squares each of side $1\frac{1}{4}$ in. (the area shaded in Fig. 34), so that $R = 16 \times \frac{25}{16} = 25$ in.²

Then the probability required is $25 / (\frac{7\frac{1}{4}}{\frac{3}{8}})^2 = \frac{400}{841} = 0.476$, assuming that the center of the penny is equally likely to fall anywhere within D .

As another example of a continuous probability distribution, consider a well-balanced, smoothly-pivoted horizontal wheel, carrying a mark on its edge and rotating above a fixed circular scale (Fig. 35). If the wheel is spun and allowed to come to rest, the pointer may be supposed equally likely to indicate any reading on the scale from 0° through 180° to 360° . The probability that it stops somewhere between, say, 0° and 90° will then be $\frac{1}{4}$. The distribution function $F(x)$ is as shown in Fig. 36, where, as usual, $F(x)$ means the probability of a value less than x , that is, in this case, of a value between 0 and x .

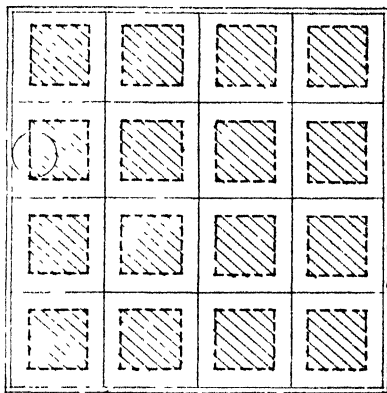


FIG. 34

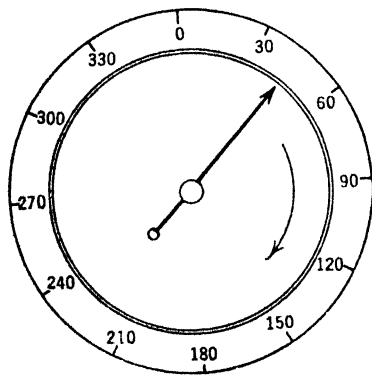


FIG. 35

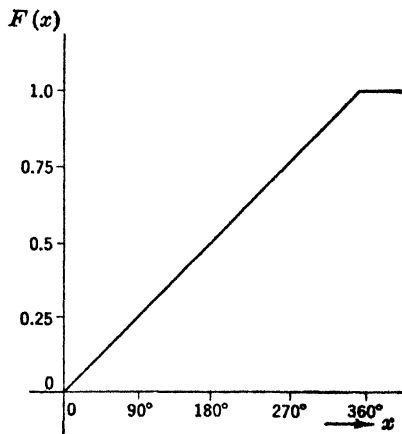


FIG. 36

The probability density $f(x)$ is constant and equal to $\frac{1}{360}$. That is to say, the probability that the pointer stops in an interval dx of the scale is $dx/360$. The probability is zero that the pointer stops precisely at some definite point of the scale, because a point is an interval of zero length. However, no point of the scale is an impossible value, because the pointer must stop somewhere. A zero probability for an event does not therefore always mean that the event is impossible.

The normal curve is an example of a continuous probability distribution in which the probability density is not constant. The probability of a value between x and $x + dx$ is $(2\pi)^{-1/2} e^{-x^2/2\sigma^2} dx$, and, as we have seen, this probability is a maximum at $x = \mu$ and falls off to zero symmetrically on both sides. The probability of a value between $x = a$ and $x = b$ is the area under the curve between these limits, and is found from the table of areas for the standard normal curve.

9.9 Moments of a Probability Distribution. We can define moments for a probability distribution in the same way as for a population, with the understanding that relative frequencies are replaced by probabilities. Thus, the mean of the distribution is given by

$$(9.21) \quad \mu = \sum_{i=1}^k x_i f(x_i)$$

if the distribution is discrete, and by

$$(9.22) \quad \mu = \int_{i_1}^{i_2} x f(x) dx$$

if the distribution is continuous. Similarly the variance is given by

$$(9.23) \quad \sigma^2 = \sum_{i=1}^k (x_i - \mu)^2 f(x_i)$$

or by

$$(9.24) \quad \sigma^2 = \int_{i_1}^{i_2} (x - \mu)^2 f(x) dx$$

according as the distribution is discrete or continuous. Higher moments can be obtained, if desired, in the same way.

Example 12. Calculate the mean and variance for the probability distribution of the number of spots with two dice (§9.7).

Here x can take integral values from 2 to 12 and the corresponding probabilities are given in the table preceding Example 9.

By adding the products of x and $f(x)$ we find from (9.21) that $\mu = 7$. (This is also obvious from the symmetry of the distribution.) From (9.23),

$$\begin{aligned} \sigma^2 &= \sum x_i^2 f(x_i) - 2\mu \sum x_i f(x_i) + \mu^2 \sum f(x_i) \\ &= \sum x_i^2 f(x_i) - \mu^2 \end{aligned}$$

since

$$\sum f(x_i) = 1 \quad \text{and} \quad \sum x_i f(x_i) = \mu$$

We find from the table that

$$\begin{aligned} \sum x_i^2 f(x_i) &= \frac{1}{36} [4 \cdot 1 + 9 \cdot 2 + 16 \cdot 3 + 25 \cdot 4 + \dots + 144 \cdot 1] \\ &= 1974/36 = 329/6. \end{aligned}$$

Therefore $\sigma^2 = 329/6 - 49 = 35/6$, so that $\sigma = 2.415$.

Example 13. Calculate the mean and variance for the probability distribution of the pointer reading in Fig. 35.

By (9.22), with $f(x) = 1/360$,

$$\mu = \frac{1}{360} \int_0^{360} x \, dx = \frac{1}{360} \cdot \frac{(360)^2}{2} = 180$$

By (9.24),

$$\sigma^2 = \int_0^{360} x^2 f(x) \, dx - \mu^2 = \frac{1}{360} \frac{(360)^3}{3} - (180)^2 = (180)^2 \left(\frac{4}{3} - 1 \right) = 10800$$

so that $\sigma = 103.9$.

9.10 Mathematical Expectation. Let A_1, A_2, \dots, A_n be mutually exclusive events, of which one must happen, and let their probabilities of occurrence be p_1, p_2, \dots, p_n . Suppose you will receive a sum of money $\$x_k$, if event A_k happens. Then we say that your mathematical expectation of gain is

$$(9.25) \quad E(x) = \sum_{k=1}^n p_k x_k$$

For example, if you buy a ticket in a lottery in which there is one prize of \$1000 and ten prizes of \$50, and if 10,000 tickets are sold, your mathematical expectation is

$$\frac{1}{10,000} (1000) + \frac{10}{10,000} (50) + \frac{9989}{10,000} (0) = \$0.15$$

since you have a probability $1/10,000$ of winning the first prize, a probability $10/10,000$ of winning one of the other prizes, and a probability $9989/10,000$ of winning nothing. Mathematical expectation is therefore different from expectation in the ordinary sense of the term, since you do not really "expect" to get 15 cents. This sum is, however, the fair price which you should pay for a ticket, in the sense that if you continued indefinitely buying tickets in similar lotteries, millions of them, your average net gain per ticket would be zero. When a man pays a premium for a term insurance policy, he is in effect playing a similar game. His beneficiary will receive the stipulated sum if he dies, and nothing if he lives. The probabilities are assessed from mortality tables, and the premium is a fair price (apart from the rather high "loading" for administration expenses, commissions, etc.).

The term "mathematical expectation", or simply *expectation*, is used in a wider sense. If $f(x_i)$ is the probability that a variate x takes the value x_i ($i = 1, 2, \dots, k$) the expectation of x is defined as

$$(9.26) \quad E(x) = \sum_{i=1}^k x_i f(x_i)$$

Similarly if x is continuous, with probability density $f(x)$, $l_1 < x < l_2$,

$$(9.27) \quad E(x) = \int_{l_1}^{l_2} x f(x) \, dx.$$

The expectation is thus the same as the mean μ for a probability distribution. Sometimes μ is called the *expected value* of x .

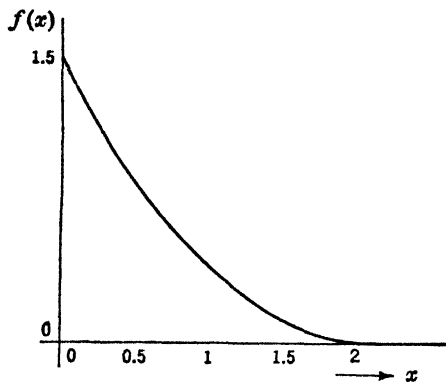


FIG. 37

Example 14. A variate x has the probability density $f(x) = 0$ for $x < 0$, $f(x) = \frac{3}{8}(x-2)^2$ for $0 < x < 2$, $f(x) = 0$ for $x \geq 2$. The graph of $f(x)$ is shown in Fig. 37.

Find the expected value of x and its standard deviation

$$\begin{aligned} E(x) &= \int_0^2 xf(x)dx \\ &= \frac{3}{8} \int_0^2 (x^3 - 4x^2 + 4x) dx \\ &= \frac{3}{8} \left[\frac{x^4}{4} - \frac{4x^3}{3} + \frac{4x^2}{2} \right]_0^2 \\ &= \frac{3}{8} \left(4 - \frac{32}{3} + 8 \right) = \frac{1}{2} \end{aligned}$$

The expectation of x^2 is similarly defined as

$$\begin{aligned} E(x^2) &= \int_0^2 x^2 f(x) dx \\ &= \frac{3}{8} \int_0^2 (x^4 - 4x^3 + 4x^2) dx \\ &= \frac{3}{8} \left(\frac{32}{5} - 16 + \frac{32}{3} \right) = \frac{3}{8} \cdot \frac{16}{15} = \frac{2}{5}. \end{aligned}$$

The variance of $x = E(x^2) - [E(x)]^2$

$$= \frac{2}{5} - \frac{1}{4} = \frac{3}{20} = 0.15$$

so that the standard deviation of x is $(0.15)^{1/2} = 0.39$.

9.11 Statistics and Probability. The theory of probability, as outlined in this chapter, forms the basis for statistical inference which will occupy us in some later chapters. The laws of probability are exact, theoretical laws: the extent to which they apply to events in the real world can be decided only by experience.

The various kinds of empirical data which form the subject matter of statistics have one element in common, namely, an element of randomness or unpredictability. Although we may feel sure that the tossing of a coin is a mechanical process governed by known physical laws, yet the result (head or tail) is in practice quite unpredictable, at least with an ordinary coin spun in the ordinary way. The final state is so dependent on minute changes in the initial position and angular velocity of the coin that we cannot possibly calculate what it will be. Similarly, the yield of corn in a given year from a given plot of land is dependent on a multitude of variable factors affecting the climate, quality of seed, etc., so that the exact value of the yield cannot be

known beforehand. The fluctuations from year to year in the date of Easter might seem to a person ignorant of Church history an example of a random process, but there exists, of course, an exact mathematical formula from which the date in any year can be predicted. The date of Easter is not, therefore, a random variable of the kind contemplated in statistics.

The bridge between the theory of probability and the behavior of statistical data is provided by the empirical fact that in a long series of random experiments or observations the relative frequency of a particular result shows a marked tendency to settle down to a constant value. In spite of the irregularities of individual tosses of a coin, and even occasional long runs of heads or tails, the proportion of heads to the total number of tosses always approximates, as this number increases, to a fixed value which is taken as the probability of head with the particular coin used. In general we can regard it as an axiom based on experience that the relative frequencies of any particular observed results in a long series of random experiments, performed under uniform conditions, will show this same kind of long-run stability and may therefore be replaced approximately by probabilities.

In practice, the long series of random experiments is often hypothetical. One or two experiments only are actually made, but these are regarded as samples from a very large, possibly even infinite, set of experiments that might conceivably be carried out, given unlimited time, opportunity, and money. The probabilities which are deduced relate to this hypothetical population.

9.12 Mathematical Models. When a set of statistical data exhibits some definite regularities we may be able to form a mathematical model of the process from which further consequences may be deduced. If, for example, a set of tosses of a coin shows a proportion of heads approximately one-half, we can replace the actual sequence of tosses, for mathematical purposes, by a random variable x which can take on the two values $x = 0$ and $x = 1$, each with probability $\frac{1}{2}$. This enables us to deduce the variance and other moments of the distribution.

In Chapter VIII we saw that the observed distribution of weights of 1000 Glasgow schoolgirls could be well fitted by a normal curve. In the process of fitting we replaced the irregular actual distribution by a symmetrical mathematical model, according to which the probability that a schoolgirl belonging to the population sampled would have a weight between x and $x + dx$ pounds is given by

$$p(x) dx = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2} dx$$

μ and σ being constants. This model is convenient for making further inferences about the population.

We shall have many examples in later chapters of the process of setting up a mathematical model. A variable x may, for instance, be suspected of increas-

ing on the average steadily with the time t (at least, over a certain interval of time) and of being also subject to random fluctuations. This dependence on t may be expressed by the mathematical model

$$x = a + bt + \epsilon, \quad \epsilon = N(0, \sigma^2)$$

which means that, apart from the random component ϵ , x is a linear function of t , and that the random component itself is normally distributed with mean zero and variance σ^2 . (See end of §8.3.) The assumption of normality is not necessary, but it is a great convenience mathematically. There is a danger in using this, or any other, mathematical model, that the assumptions made may not actually be satisfied by the data supposed to be represented by the model. If there is grave doubt about this, the model may have to be changed.

Exercises

1. If A and B are independent events, with $P\{A\} = \frac{1}{3}$ and $P\{B\} = \frac{2}{3}$, what is $P\{A + B\}$? *Ans.* $\frac{5}{9}$. *Hint.* $P\{AB\} = P\{A\}P\{B\}$.

2. If $P\{A\} = \frac{1}{3}$, $P\{B\} = \frac{2}{3}$, and $P\{A + B\} = \frac{1}{2}$, what is $P\{B|A\}$? *Ans.* $\frac{1}{2}$.

3. Prove that if $P\{B|A\} = P\{B\}$, then also $P\{B|\bar{A}\} = P\{B\}$, $P\{A|B\} = P\{A\}$ and $P\{A|\bar{B}\} = P\{A\}$.

4. If A, B, C are three events, prove that
 $P\{A + B + C\} = P\{A\} + P\{B\} + P\{C\} - P\{AB\} - P\{AC\} - P\{BC\} + P\{ABC\}$.

Hint. Denote the event $B + C$ by D . Use (9.9) for $P\{A + D\}$ and then for $P\{D\}$. Note that $P\{AD\} = P\{AB + AC\} = P\{AB\} + P\{AC\} - P\{ABC\}$.

5. (a) How many 5-digit numbers are there with every digit odd? (b) How many are there with no digit lower than 6? *Ans.* (a) 3125; (b) 1024.

6. How many numbers greater than a million can be formed from the digits 2, 3, 0, 3, 4, 2, 3? *Ans.* 360.

7. How many arrangements can be made of the letters of the word "draught," if the vowels are never separated? *Ans.* 1440.

8. (a) How many "words" (i.e., arrangements of letters) can be formed using all the letters of the word "article"? (b) In how many of these do vowels occupy the even places? *Ans.* (a) 5040; (b) 144.

9. If $4P(n, 3) = 5P(n - 1, 3)$, what is n ?

10. At a dinner table the host and hostess sit opposite each other. In how many ways can $2n$ guests be arranged so that two particular guests do not sit together?

Ans. $2(2n^2 - 3n + 2)(2n - 2)!$

11. Four strangers board a bus, in which there are 6 empty seats. In how many different ways can they be seated?

12. Six examination papers are set in a comprehensive examination, two of them in mathematics. In how many different orders can the papers be given if the two mathematics papers are not to be successive? *Ans.* 480.

13. An 8-oared boat is to be manned by a crew chosen from 11 men. In how many ways can the crew be chosen and arranged if 3 men can steer but cannot row, and the rest can row but cannot steer, and if 2 men can row only on the port side? *Ans.* 25,920.

14. Show that the number of ways in which p positive and n negative signs may be placed in a row so that no two negative signs shall be together is $C(p + 1, n)$, for all nonnegative integral values of p and n .

Hint. Use Theorem 7. The n negative signs can be thought of as partitions between $n + 1$ compartments in which $+$ signs are distributed. $n - 1$ of these must be placed one in each compartment, except the end ones. The rest can be distributed at pleasure.

15. Prove that $rC(n, r) = nC(n - 1, r - 1)$.

16. Prove that $\sum_{r=1}^n rC(n, r) = 2^{n-1}n$.

Hint. Use exercise 15. $\sum_{r=1}^n C(n - 1, r - 1) = \sum_{r=0}^{n-1} C(n - 1, r)$, which is the total number of ways of selecting 0 or 1 or 2 \dots or $n - 1$ out of $n - 1$ different things.

17. Six cards are drawn at random from a deck of 52 cards. What is the probability that 3 will be red and 3 black? *Ans.* 0.332.

18. If 4 cards are drawn at random from a deck of 52 cards, what is the probability that there will be one card of each suit? *Ans.* 0.1055

19. If 30 similar balls are placed at random in 8 bags, empty bags being admissible, what is the probability that no bag contains less than 3 balls? *Ans.* 13/77,996.

20. A room has 3 lamp sockets. From a collection of 10 light bulbs, of which 6 are no good, a person selects 3 at random and puts them in the sockets. What is the probability that he will have light? *Ans.* $\frac{5}{8}$.

Hint. Find the probability of *not* getting light, i.e., of selecting 3 bad bulbs.

21. If $C(n, 12) = C(n, 8)$, what is n ? What is $C(22, n)$?

22. If a man has 6 friends, in how many ways can he invite one or more of them to dinner? *Ans.* 63.

23. An urn contains 12 balls of which 3 are marked. If 5 balls are drawn out together, what is the probability that all three of the marked balls are among these 5? *Ans.* $\frac{1}{11}$.

24. If p is the probability of occurrence of an event in a single trial, show that the probability of at least one occurrence in n independent trials is $1 - (1 - p)^n$.

25. What is the chance of throwing 6 with a die at least once in 5 trials? *Ans.* 0.598.

26. What is the chance that a hand at bridge contains the Ace and King of Spades? *Ans.* $\frac{1}{17}$.

27. A batch of 1000 lamps is known to have 5% defectives. If 5 lamps chosen at random are tested, what is the probability that none of them will be defective? What is the probability that exactly 2 defectives will be found?

Ans. (a) $C(950, 5)/C(1000, 5)$; (b) $C(50, 2)C(950, 3)/C(1000, 5)$.

28. A manufacturer supplies cheap clocks in lots of 50. A buyer, before taking a lot, tests a random sample of 5 clocks, and if all are good he accepts the lot. Otherwise he refuses it. What is the probability that he will accept a lot containing 10 defective clocks? What is the probability that he will reject a lot containing only one defective clock?

Ans. (a) 0.31; (b) 0.1.

29. A factory produces a certain type of screw, put up in boxes of 100. Boxes are inspected by taking 20 screws at random out of the box and rejecting the box if any defectives are found. What is the probability of passing a box containing 2 defective screws?

Ans. 0.638.

30. An enemy factory covers 2.5 acres, and the power plant in this factory occupies 200 square yards. If a bomb is dropped on the factory from a high altitude, it may be supposed equally likely to strike anywhere. What is the probability that, if a bomb does hit the factory, it hits the power plant? How many bombs must be dropped to give a probability of over 0.9 that at least one will hit the power plant (1 acre = 4840 sq yd)?

Ans. (a) 2/121; (b) 139 at least.

31. Two points are marked at random on a straight line of length a . What is the probability that the distance between them will exceed c , where $c < a$? *Ans.* $(a - c)^2/a^2$.

Hint. If the first point is not within a distance c of either end of the line, the second

point can be anywhere except in the interval $2c$ centered on the first point. The probability of the combined event is $(a - 2c)^2/a^2$. If the first point is within distance x of either end ($x < c$), the excluded interval is $c + x$. The probability of the joint event for any $x < c$ is

$$\int_0^c \frac{2dx}{a} \cdot \frac{a - c - x}{a}.$$

32. Calculate the mean and variance and sketch the graph of the following *rectangular distribution*:

$$\begin{cases} f(x) = \frac{1}{2}, & -1 < x < 1 \\ f(x) = 0 & \text{for } x < -1 \text{ and for } x > 1 \end{cases}$$

Sketch the distribution function $F(x)$.

33. Calculate the mean and variance and sketch the graph of the following *triangular distribution*:

$$\begin{cases} f(x) = 2(x + \sqrt{2}/2), & -\sqrt{2}/2 \leq x < 0 \\ f(x) = 2(\sqrt{2}/2 - x), & 0 \leq x \leq \sqrt{2}/2 \\ f(x) = 0, & x < -\sqrt{2}/2 \text{ or } x > \sqrt{2}/2 \end{cases}$$

Ans. 0, $\frac{1}{\sqrt{2}}$.

Hint. Integrate separately for the two regions — $\sqrt{2}/2$ to 0 and 0 to $\sqrt{2}/2$.

34. The probability density of a variate x is

$$\begin{cases} f(x) = 0, & x < 1 \\ f(x) = \frac{-3x^3}{8} + \frac{3x^2}{2} - \frac{9x}{8}, & 1 \leq x \leq 3 \\ f(x) = 0, & x > 3 \end{cases}$$

(a) Verify that the area under the curve is unity. (b) Sketch the graph. (c) Find the mean and standard deviation of x . (d) Find the probability that $1 < x < 1\frac{1}{2}$.

Ans. (c) 2.1, 0.436; (d) 53/512.

35. A continuous distribution of a variate x is defined by

$$\begin{cases} f(x) = \frac{x}{2}, & 0 \leq x \leq 1 \\ f(x) = \frac{1}{2}, & 1 \leq x \leq 2 \\ f(x) = \frac{1}{2}(3 - x), & 2 \leq x \leq 3 \end{cases}$$

Sketch the distribution and find the variance of x . *Ans.* $\frac{5}{12}$.

36. Two defective radio tubes were accidentally placed in a box with 5 nondefective tubes. The tubes are tested one at a time until the second defective is found. Compute the probabilities $f(x)$ that the x th tube tested is the second defective. Find the mean and variance of x . *Ans.* $5\frac{1}{3}$, $2\frac{2}{3}$.

Hint. $x = 2, 3, 4, 5, 6$, or 7 . Find the probability that exactly one tube is defective in the first $x - 1$ tested and that then the next tube tested is also defective. Evaluate for the different values of x .

37. A bag contains 5 nickels and a quarter, all being wrapped in paper, so as to be indistinguishable. A boy is allowed to draw one coin at a time and keep it until he draws the quarter, when he must stop. What is his expectation? *Ans.* 37.5 cents.

38. A throws 6 pennies on the table, and pays B 6 dollars if either 6 heads or 6 tails appear, and 5 dollars if 5 heads or 5 tails appear. In every other case he takes B's stake. How much should this stake be to make the game fair? *Ans.* \$1.44.

39. There are three identical-appearing envelopes in a drawer. One contains two \$1 bills and one \$10 bill, the second contains one \$1 bill and two \$10 bills, the third contains

three \$1 bills. If a man is allowed to pick one envelope and draw one bill from the envelope without looking at it, what is his expectation? *Ans.* \$4.

Hint. The conditions are equivalent to picking one bill at random from 3 tens and 6 ones.

40. A circle of diameter 8 inches is drawn in the interior of a square of side 12 in. A penny (diameter $\frac{3}{4}$ in.) is dropped on the square, which is lying on a horizontal table. If only those cases are counted when the penny lies completely inside the square, what is the probability that at least part of the coin lies outside the circle? *Ans.* 0.674.

41. The floor of a large room is made of hardwood, laid in strips one inch wide, with cracks between of negligible width. A coin of diameter $1\frac{1}{2}$ in. is dropped on the floor. Find the probability that the coin touches three strips.

References

1. For a discussion of the frequency definition of probability by one of its most prominent exponents, see R. von Mises, *Probability, Statistics, and Truth* (W. Hodge, 1939).

2. A popular account of probability as applied to games of chance and advertising, is H. C. Levinson, *The Science of Chance, from Probability to Statistics* (Rinehart & Company, Inc., 1950).

3. The student interested in the more serious side of probability will find a good account of applications to genetics in J. Neyman, *First Course in Probability and Statistics* (Henry Holt & Co., 1950).

4. An excellent account of probability for the more mature student is W. Feller, *Introduction to Probability Theory and its Applications*, Vol. I (John Wiley & Sons, Inc., 1950). Vol. I deals with discrete distributions only.

5. A more popular work, profusely illustrated by red and black diagrams, is L. Hogben, *Chance and Choice by Card Pack and Chessboard* (Chanticleer Press, Inc., 1950).

CHAPTER X

THE BINOMIAL AND POISSON DISTRIBUTIONS

10.1 A Coin-tossing Problem. Suppose we toss a coin s times and ask what is the probability of getting exactly x heads. We can symbolize the results of a set of tosses by a row of 0's and 1's, in which a 0 means "head" and a 1 means "tail". Thus,

0 0 1 0 1 1 1 1 0 1 0 0 1 1 0 0 0 1 0 1 . . .

We will suppose that, for this coin, head and tail are equally likely to fall uppermost, so that the probability of head on a single toss is $\frac{1}{2}$. We do not care what the arrangement of 0's and 1's may be, as long as there are just x 0's and $s - x$ 1's. The number of favorable arrangements is therefore the number of permutations of s things, x alike of one kind and $s - x$ alike of another kind, and by Theorem 3 (Corollary) of Chapter IX this is $s!/[x!(s - x)!]$. The total number of ways in which the tosses may turn out is 2^s , since in each of the s independent tosses there are 2 possibilities. The probability required is therefore

$$(10.1) \quad f(x) = \frac{s!}{x!(s - x)!} \left(\frac{1}{2}\right)^s = C(s, x) \left(\frac{1}{2}\right)^s$$

This is a discrete distribution, since x is obviously an integer between 0 and s inclusive. Thus, for $s = 4$, x can take the values 0, 1, 2, 3, or 4, with probabilities given by

$$\begin{array}{c|cccccc} x & 0 & 1 & 2 & 3 & 4 \\ \hline 16f(x) & 1 & 4 & 6 & 4 & 1 \end{array}$$

The probability of exactly 3 heads in 4 tosses is therefore $\frac{4}{16} = \frac{1}{4}$. The probability of *at least* 2 heads is $(6 + 4 + 1)/16 = \frac{11}{16}$, and so on.

10.2 Binomial Coefficients. The quantities $C(s, x)$ which appear in (10.1) are called binomial coefficients because they can be obtained by expanding the binomial expression $q + p$ raised to the s th power. Thus

$$(10.2) \quad \begin{cases} (q + p)^2 = q^2 + 2qp + p^2 \\ (q + p)^3 = q^3 + 3q^2p + 3qp^2 + p^3 \\ (q + p)^4 = q^4 + 4q^3p + 6q^2p^2 + 4qp^3 + p^4 \\ \text{etc.,} \end{cases}$$

The coefficients in the first line of (10.2) are $C(2, 0)$, $C(2, 1)$, and $C(2, 2)$;

to its immediate left and right. Thus in the 5th line ($s = 4$), we have $4 = 1 + 3$, $6 = 3 + 3$, etc.

10.3 The Binomial Distribution. Instead of the tossing of a coin let us consider an event A of which the probability in a single trial is p . We suppose that we can make a succession of independent trials, and that in each one the event A either happens or does not happen. If, for example, A is the throwing of a six with a good die, \bar{A} is the event of *not* throwing a six (that is, of throwing any other number), and since either A or \bar{A} must happen, $P\{A\} + P\{\bar{A}\} = 1$. We shall denote the probability of \bar{A} by q , and therefore

$$(10.5) \quad p + q = 1$$

We require the probability of exactly x successes in s trials, a success being the event " A happens." Since the trials are independent, the product law of probability holds, and the probability of any given succession of A 's and \bar{A} 's, such as

$$A \ A \ \bar{A} \ \bar{A} \ \bar{A} \ \bar{A} \ A \ \bar{A} \ A \ A \ \bar{A} \ A \ \bar{A} \ \bar{A} \cdots$$

is

$$p \ p \ q \ q \ q \ q \ p \ q \ p \ p \ q \ p \ q \ q \cdots$$

If there are x A 's and $(s - x)$ \bar{A} 's in this sequence, the probability is $p^x q^{s-x}$. But if we are interested only in the *total number* of A 's and \bar{A} 's, and do not care where they come in the sequence, we must multiply this probability by the number of ways of permuting x A 's and $(s - x)$ \bar{A} 's, all ways being supposed equally likely. This number is $C(s, x)$, and therefore the required probability is

$$(10.6) \quad f(x) = C(s, x) p^x q^{s-x} = \frac{s!}{x!(s-x)!} p^x q^{s-x}$$

which, as we have seen, is precisely the general term in the expansion of $(q + p)^s$. The distribution given by (10.6), for values of $x = 0, 1, 2, \dots, s$, is called the *binomial distribution*, or the *Bernoulli distribution*, after James Bernoulli (1654–1705) who stated the foregoing result in his posthumously published *Ars Conjectandi* in 1713.

In order to calculate successive values of $f(x)$, it is usually convenient to use a *recursion formula*, that is, a formula which gives $f(x + 1)$ when $f(x)$ is known. Thus by starting with $f(0)$ we can form successively $f(1)$, $f(2)$, and so on. The formula is

$$(10.7) \quad f(x + 1) = \frac{s - x}{x + 1} \cdot \frac{p}{q} f(x)$$

Proof: By (10.6),

$$f(x + 1) = \frac{s!}{(x + 1)!(s - x - 1)!} p^{x+1} q^{s-x-1}$$

Therefore, dividing $f(x + 1)$ by $f(x)$, we have

$$\frac{f(x + 1)}{f(x)} = \frac{x!(s - x)!}{(x + 1)!(s - x - 1)!} \frac{p^{x+1} q^{s-x-1}}{p^x q^{s-x}} = \frac{s - x}{x + 1} \frac{p}{q}$$

Now

$$f(0) = p^0 q^s = q^s, \quad f(1) = \frac{s-0}{0+1} \frac{p}{q} f(0) = \frac{s}{1} \frac{p}{q} f(0)$$

$$f(2) = \frac{s-1}{2} \frac{p}{q} f(1)$$

and so on.

If, for example, $s = 5$ and $p = \frac{1}{6}$, we can calculate the respective values of $f(x)$ for $x = 0, 1, 2, 3, 4, 5$, as follows:

$$f(0) = \left(\frac{5}{6}\right)^5 = 0.4019$$

$$f(1) = 5 \cdot \frac{1}{5} f(0) = 0.4019$$

$$f(2) = \frac{4}{2} \cdot \frac{1}{5} f(1) = 0.4 f(1) = 0.1608$$

$$f(3) = \frac{3}{3} \cdot \frac{1}{5} f(2) = 0.2 f(2) = 0.0322$$

$$f(4) = \frac{2}{4} \cdot \frac{1}{5} f(3) = 0.1 f(3) = 0.0032$$

$$f(5) = \frac{1}{5} \cdot \frac{1}{5} f(4) = 0.04 f(4) = 0.0001$$

Apart from a slight error of rounding-off, these add up to 1, as they should.

The binomial distribution may be represented by a histogram. If we construct rectangles of unit base, centered at $x = 0, 1, 2, \dots, s$, with heights equal to $f(x)$, the total area of the histogram will be unity. The histogram for $s = 5$ and $p = \frac{1}{6}$ is shown in Fig. 38.

Notice that in this distribution all the values of x are actually concentrated at the centers of the intervals, so that there is no grouping error. The distribution function is stepped, like that in Fig. 8 of §2.6.

The modal value of x in the binomial distribution may be found as follows:

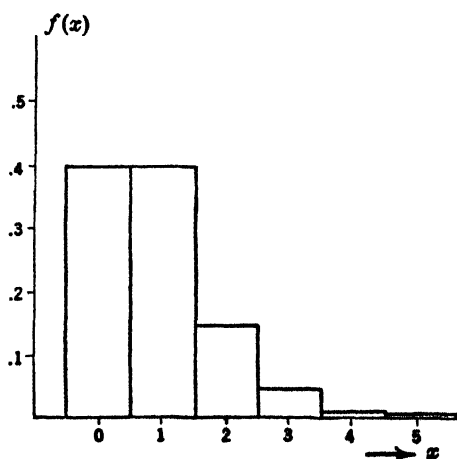


Fig. 38

If \hat{x} is the mode, $f(\hat{x})$ must be at least as great as the frequencies for the adjacent values $\hat{x} - 1$ and $\hat{x} + 1$, that is,

$$\frac{f(\hat{x})}{f(\hat{x} - 1)} \geq 1 \quad \text{and} \quad \frac{f(\hat{x})}{f(\hat{x} + 1)} \geq 1$$

Substituting the appropriate values for $f(\hat{x})$, $f(\hat{x} + 1)$ and $f(\hat{x} - 1)$ from (10.6), we find

$$\frac{f(\hat{x})}{f(\hat{x} - 1)} = \frac{s!}{\hat{x}!(s - \hat{x})!} \frac{(\hat{x} - 1)!(s - \hat{x} + 1)!}{s!} \frac{p^{\hat{x}} q^{s - \hat{x}}}{p^{\hat{x} - 1} q^{s - \hat{x} + 1}} = \frac{s - \hat{x} + 1}{\hat{x}} \frac{p}{q}$$

and similarly

$$\frac{f(\hat{x})}{f(\hat{x} + 1)} = \frac{\hat{x} + 1}{s - \hat{x}} \frac{q}{p}$$

Hence we see that \hat{x} satisfies simultaneously the relations:

$$\frac{s - \hat{x} + 1}{\hat{x}} \frac{p}{q} \geq 1, \quad \frac{\hat{x} + 1}{s - \hat{x}} \frac{q}{p} \geq 1$$

These can be written

$$(s + 1)p \geq (p + q)\hat{x}, \quad sp - q \leq (p + q)\hat{x}$$

or, since $p + q = 1$,

$$(10.8) \quad \hat{x} \leq sp + p, \quad \hat{x} \geq sp + p - 1$$

Since \hat{x} is an integer, \hat{x} is uniquely determined by the equations (10.8), unless $(s + 1)p$ happens also to be an integer. If so, the two values $sp + p$ and $sp + p - 1$ correspond to equal values of $f(\hat{x})$ and the mode is not unique. This is the situation in Fig. 38, where $(s + 1)p = 1$, and $f(0) = f(1)$.

Example 1. What is the most probable number of times that an ace will appear if a die is tossed (a) 50 times (b) 53 times?

Assuming that the probability of ace is $\frac{1}{6}$, we have (a) $s = 50$, $p = \frac{1}{6}$, $\hat{x} \leq 8.5$, and $\hat{x} \geq 7.5$, so that the most probable value is $\hat{x} = 8$; (b) $s = 53$, $p = \frac{1}{6}$, $\hat{x} \leq 9$, and $\hat{x} \geq 8$, so that now the numbers 8 and 9 are equally probable.

10.4 Moments of the Binomial Distribution. The binomial distribution for assumed values of p and s is a theoretical discrete distribution with p and s as parameters, so that according to the convention of distinguishing parameters by Greek letters, we should write, say, θ and ν instead of p and s . (The symbol π is sometimes used, as the Greek form of p , but there is a risk of confusion with the customary meaning of π .) We shall deal in the next chapter with the problem of samples from a population in which the variate is assumed to be binomial, with a probability of success which is estimated from the relative frequency of success in the sample, and it will then be necessary to distinguish between the relative frequency p , and the probability θ ,

of which p is an estimate. In the present section, therefore, we will rewrite the binomial distribution law as

$$(10.9) \quad f(x) = C(s, x)\theta^x(1 - \theta)^{s-x}$$

which brings out the fact that θ is a parameter. It is true that s is also a parameter, but in most practical applications s is an integer determined by the nature of the problem and is not estimated from the sample.

By the usual formulas for a discrete distribution, the mean (or expectation) of x is given by

$$(10.10) \quad \mu = \sum_{i=0}^s x_i f(x_i)$$

and the variance by

$$(10.11) \quad \sigma^2 = \sum_{i=0}^s x_i^2 f(x_i) - \mu^2$$

For the binomial distribution,

$$x_0 = 0, x_1 = 1, x_2 = 2, \dots, x_s = s$$

so that

$$(10.12) \quad \begin{aligned} \mu &= \sum_{k=0}^s k \frac{s!}{k!(s-k)!} \theta^k (1 - \theta)^{s-k} \\ &= \sum_{k=1}^s \frac{s!}{(k-1)!(s-k)!} \theta^k (1 - \theta)^{s-k} \end{aligned}$$

since the term with $k = 0$ vanishes. Also,

$$(10.13) \quad \sigma^2 = \sum_{k=1}^s k^2 \frac{s!}{k!(s-k)!} \theta^k (1 - \theta)^{s-k} - \mu^2$$

Now, by the binomial theorem,

$$(10.14) \quad \begin{aligned} \{\theta + (1 - \theta)\}^s &= \sum_{k=0}^s C(s, k) \theta^k (1 - \theta)^{s-k} \\ &= 1 \end{aligned}$$

since

$$\theta + (1 - \theta) = 1$$

From (10.12), taking s as a factor out of $s!$ and θ out of θ^k , we have

$$\begin{aligned} \mu &= \sum_{k=1}^s \frac{s(s-1)! \theta \cdot \theta^{k-1} (1 - \theta)^{s-k}}{(k-1)!(s-k)!} \\ &= s\theta \sum_{k=1}^s \frac{(s-1)! \theta^{k-1} (1 - \theta)^{s-k}}{(k-1)!(s-k)!} \\ &= s\theta \sum_{k-1=0}^{s-1} C(s-1, k-1) \theta^{k-1} (1 - \theta)^{s-1-(k-1)} \\ &= s\theta \end{aligned}$$

by (10.14), with $k - 1$ instead of k and $s - 1$ instead of s . We obtain, therefore, the important relation

$$(10.15) \quad \mu = s\theta$$

To calculate σ^2 from (10.13) it is convenient to write $k^2 = k(k - 1) + k$, since k and $k - 1$ are factors of $k!$. Then

$$\sigma^2 = \sum_{k=0}^s \frac{\{k(k-1) + k\}}{k!(s-k)!} s! \theta^k (1-\theta)^{s-k} - \mu^2$$

The second term in the bracket, k , gives precisely μ . The first term gives, on canceling $k(k-1)$ into $k!$,

$$\begin{aligned} \sum_{k=2}^s \frac{s! \theta^k (1-\theta)^{s-k}}{(k-2)!(s-k)!} &= s(s-1)\theta^2 \sum_{k=2}^{s-2} \frac{(s-2)! \theta^{k-2} (1-\theta)^{s-k}}{(k-2)!(s-k)!} \\ &= s(s-1)\theta^2 \end{aligned}$$

by (10.14), with $s - 2$ and $k - 2$ instead of s and k .

Therefore

$$\begin{aligned} \sigma^2 &= s(s-1)\theta^2 + \mu - \mu^2 \\ &= s(s-1)\theta^2 + s\theta - s^2\theta^2 \\ &= s(\theta - \theta^2) \\ (10.16) \quad &= s\theta(1-\theta) \end{aligned}$$

The standard deviation of the binomial distribution is therefore $[s\theta(1-\theta)]^{1/2}$.

Higher moments can be calculated by the same method, but can be more simply obtained by a recursion formula. (See Part Two, §2.8.) Here we simply mention that the third moment is

$$(10.17) \quad \mu_3 = s\theta(1-\theta)(1-2\theta)$$

whence the moment measure of skewness is

$$\begin{aligned} (10.18) \quad \alpha_3 &= (1-2\theta)/[s\theta(1-\theta)]^{1/2} \\ &= (1-2\theta)/\sigma \end{aligned}$$

The binomial distribution is therefore always skew unless $\theta = \frac{1}{2}$.

10.5 Fitting a Binomial to a Given Distribution. A simple experiment in sampling may be conducted by a class as follows: A mixture of 100 red and 200 white balls (identical except for color) is placed in a box. These may be wooden balls, about 1 cm in diameter. Samples of 10 balls are extracted by stirring up the contents and inserting a small wooden paddle, in the upper side of which are 10 holes, into which the balls will fit easily. By dipping the paddle into the mixture and bringing it up, a sample of 10 is easily examined. The number of red balls in the sample is noted, and the balls are then returned to the box, and the procedure is repeated. It does not take long for

a class of students to compile three or four hundred samples in this way, and the results can then be set out as a frequency distribution. A distribution so obtained is shown in Table 35, columns 1 and 2, where x is the number of red balls in the sample and f_0 the observed frequency of x .

TABLE 35. BINOMIAL DISTRIBUTION FITTED TO SAMPLING RESULTS

x	f_0	$f_c(\theta = \frac{1}{3})$	$f_c(\theta = 0.36)$	xf_0	x^2f_0
0	2	6.1	4.0	0	0
1	22	30.3	22.7	22	22
2	63	68.3	57.5	126	252
3	76	91.0	86.2	228	684
4	96	79.7	84.8	384	1536
5	56	47.8	57.3	280	1400
6	26	19.9	26.8	156	936
7	8	5.7	8.6	56	392
8	1	1.1	1.8	8	64
9	0	0.1	0.2	0	0
10	0	0.0	0.0	0	0
	350	350.0	349.9	1260	5286

This method of sampling does not correspond exactly to a binomial situation, since the 10 balls are taken out at once. Strictly, the balls should be picked one at a time and each one returned to the box after its color has been noted, but this takes too long and the method described above is a good approximation.

We can now compare the distribution obtained with a theoretical binomial distribution. Since exactly one-third of the balls in the box are red, and since every ball has an equal chance (approximately) of being included in the sample, we can take the theoretical probability θ as $\frac{1}{3}$. The probability of x red balls in a sample of 10 is given by

$$(10.19) \quad f(x) = C(10, x) \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{10-x}$$

so that the calculated frequency f_c in 350 samples is

$$(10.20) \quad f_c = 350 f(x)$$

The values of f_c are given in column 3 of Table 35. Thus,

$$f_c(0) = 350 \left(\frac{2}{3}\right)^{10} = 6.07$$

$$f_c(1) = \frac{10}{1} \cdot \frac{1}{2} f_c(0) = 30.35, \text{ by (10.7)}$$

$$f_c(2) = \frac{9}{2} \cdot \frac{1}{2} f_c(1) = 68.28$$

etc.

The observed and calculated frequencies correspond roughly, but the agreement does not look to be remarkably good. A method of judging how good it is (the "chi-square" test) will be given in Chapter 13. Meanwhile, we can calculate the mean and variance of the sample distribution and compare them with the theoretical values.

The distribution mean \bar{x} is found by calculating $\sum xf_0 = 1260$ (see column 5 of Table 35). Then $\bar{x} = \frac{1260}{350} = 3.60$, and this may be taken as the estimate of μ for the population consisting of all possible samples of 10 from the collection of red and white balls. Actually 350 samples were obtained, but these form only a very small fraction of all the samples that could be taken with unlimited time and patience. The distribution variance s_x^2 is given by

$$s_x^2 = \frac{5286}{350} - \left(\frac{1260}{350} \right)^2 = 15.103 - 12.96 = 2.143$$

and an unbiased estimate of σ^2 is $\tau^2 = 350 s_x^2 / 349 = 2.149$.

The theoretical values of μ and σ^2 are $s\theta = 3.33$ and $s\theta(1 - \theta) = 2.22$, respectively.

Without assuming the value $\frac{1}{3}$ for θ , we can fit a binomial to our frequency distribution in the same way as we fitted a normal curve to a given distribution, namely, by estimating the parameter θ from the distribution itself. That is, we take for θ the value $\bar{x}/s = 0.36$. The calculated frequencies are now given by

$$(10.21) \quad f_c = 350 C(10, x) (0.36)^x (0.64)^{10-x}$$

Then

$$f_c(0) = 350 (0.64)^{10} = 4.035$$

$$f_c(1) = \frac{10}{1} \cdot \frac{9}{16} f_c(0) = 22.70$$

etc.

The values are given in column 4 of Table 35. They agree much better than before with the observed values. The mean is, of course, now identical with the observed mean because it was made to be so. The new theoretical variance is $10 \times 0.36 \times 0.64 = 2.30$, so that the agreement in this respect is rather worse than before.

10.6 The Poisson Distribution. Situations sometimes occur where the probability θ of a certain event is very small, but where nevertheless the number of trials s is so large that the expected number of occurrences of the event is of moderate size, say between 0.1 and 10. Examples are the number of persons born blind per year in a large city, the number of typographical errors made by a good typist in a large number of typed pages, the number of bridge hands containing 4 aces in an evening of play at a bridge club. The events

here are "rare events," that is, *individually*. The probability of a blind birth is very small, but the number of persons born per year in a large city is large, so that such births taken over the whole city and the whole year are not rare. In situations like these, the true binomial expression for the probability of occurrence of x events can be replaced by a convenient approximation due to S. D. Poisson (1781-1840) and known as the Poisson distribution.

We suppose, then, that θ is small and s large, in such a way that $\lambda = s\theta$ is a number of the order of unity. We want to find an expression for the binomial probability,

$$\begin{aligned} f(x) &= \frac{s!}{x!(s-x)!} \theta^x (1-\theta)^{s-x} \\ &= \frac{s(s-1)(s-2) \cdots (s-x+1)}{x!} \left(\frac{\lambda}{s}\right)^x \left(1 - \frac{\lambda}{s}\right)^{s-x} \end{aligned}$$

The number of factors in the numerator of the first fraction is x , and if we divide each of them by s , we account for the term s^x in the denominator of $f(x)$. Therefore,

$$(10.22) \quad f(x) = \frac{1\left(1 - \frac{1}{s}\right)\left(1 - \frac{2}{s}\right) \cdots \left(1 - \frac{x-1}{s}\right) \lambda^x \left(1 - \frac{\lambda}{s}\right)^s}{x! \left(1 - \frac{\lambda}{s}\right)^x}$$

Now, for a fixed value of x , we let s become very large, tending to infinity.

Then all the fractions $\frac{1}{s}, \frac{2}{s}, \dots, \frac{x-1}{s}, \frac{\lambda}{s}$ become very small, and all the

terms like $1 - \frac{1}{s}, 1 - \frac{x-1}{s}, 1 - \frac{\lambda}{s}$ are practically equal to 1. When

$\left(1 - \frac{\lambda}{s}\right)$ is raised to a fixed power x , it is still practically 1, but in the numerator of (10.22) it is raised to the power s , which is very large and ultimately infinite. It is shown in most textbooks on elementary calculus that the limit of $\left(1 - \frac{\lambda}{s}\right)^s$ as $s \rightarrow \infty$ is $e^{-\lambda}$, where e is the number (2.71828...) which we have already encountered in the normal law and which is the base of natural logarithms. Hence, when $s \rightarrow \infty$,

$$(10.23) \quad f(x) \rightarrow \frac{\lambda^x}{x!} e^{-\lambda} = P(x, \lambda)$$

which is Poisson's function. For s large but finite, the expression on the right-hand side of (10.23) can be taken as a good *approximation* to $f(x)$. The advantages are that it is comparatively easy to calculate and that good tables of $P(x, \lambda)$ exist. (See Reference 1.)

The Poisson distribution is, like the binomial, a discrete distribution, since x must be a nonnegative integer. There is no upper limit on x , but the probabilities of large values of x are extremely small.

As in §10.3 it is easily proved that the mode of the Poisson distribution is the integer lying between $\lambda - 1$ and λ , unless λ happens to be an integer, in which case the values for λ and $\lambda - 1$ are equal.

10.7 Moments of the Poisson Distribution. If we add the values of $P(x, \lambda)$ for all values of x , we obtain

$$\sum_{x=0}^{\infty} P(x, \lambda) = e^{-\lambda} \left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots \right)$$

Now, it can be proved that the series in parentheses has a sum for any value of λ , and this sum is e^{λ} . Therefore

$$(10.24) \quad \sum_{x=0}^{\infty} P(x, \lambda) = 1$$

which is, of course, what we should expect if $P(x, \lambda)$ is actually the probability of the value x .

The expectation of x is given by

$$\begin{aligned} \mu &= \sum_{x=0}^{\infty} x P(x, \lambda) \\ &= \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\ &= e^{-\lambda} \left[\lambda + \frac{\lambda^2}{1!} + \frac{\lambda^3}{2!} + \cdots \right] \\ (10.25) \quad &= \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda \end{aligned}$$

The expected value of x is therefore λ or $s\theta$, as with the binomial law.

The variance of x is

$$\begin{aligned} \sigma^2 &= \sum_{x=0}^{\infty} x^2 P(x, \lambda) - \mu^2 \\ &= \sum_{x=1}^{\infty} \frac{\{x(x-1) + x\} \lambda^x e^{-\lambda}}{x!} - \lambda^2 \\ &= \lambda \sum_{x=1}^{\infty} \frac{(x-1) + 1}{(x-1)!} \lambda^{x-1} e^{-\lambda} - \lambda^2 \\ &= \lambda \sum_{x=0}^{\infty} \frac{x+1}{x!} \lambda^x e^{-\lambda} - \lambda^2 = \lambda(\lambda+1) - \lambda^2 = \lambda \end{aligned}$$

by (10.24) and (10.25).

The variance of x is therefore also equal to λ . This fact, that the variance is equal to the mean, is a characteristic of the Poisson distribution.

Higher moments can be found in the same way. It turns out that the third moment μ_3 is again equal to λ , so that the skewness α_3 is $\lambda^{-1/2}$. The kurtosis γ_2 ($= \alpha_4 - 3$) is λ^{-1} . For large values of λ , the skewness and kurtosis are both nearly zero, and the distribution can then be closely fitted by a normal curve.

10.8 Fitting a Poisson Distribution to a Given Empirical Distribution.

A sampling experiment similar to the binomial one described in §10.5 can be carried out with colored balls, using, say, a mixture of 100 red balls and 1100 white ones, and taking samples of 50 with a larger paddle. The probability of picking a red ball is about $\frac{1}{12}$, and therefore $\lambda = \frac{50}{12} = 4.17$. Here we have neither a very small θ nor a very large s , but the experiment does show that the Poisson approximation is reasonably good. The results of one such classroom experiment are shown in Table 36. No observed value of x in the 300 samples taken was larger than 9, but the theoretical values continue for $x = 10, 11, 12 \dots$, and so the values for all x greater than or equal to 9 are lumped together for convenience.

The distribution mean is $\bar{x} = 1324/300 = 4.413$, which is an estimate of λ , and is somewhat higher than the theoretical value 4.17 based on the assumption that all the balls are equally likely to be picked. The distribution variance is $s_x^2 = 3.756$, and the estimated σ^2 is therefore 3.768, somewhat lower than the theoretical λ .

TABLE 36. FITTING OF POISSON DISTRIBUTION

x	f_0	$f_c(\lambda = 4.17)$	$f_c(\lambda = 4.41)$	xf_0	x^2f_0
0	1	4.7	3.6	0	0
1	16	19.4	16.0	16	16
2	36	40.4	35.4	72	144
3	48	56.1	52.1	144	432
4	62	58.4	57.5	248	992
5	51	48.7	50.7	255	1275
6	41	33.8	37.3	246	1476
7	22	20.1	23.5	154	1078
8	18	10.5	13.0	144	1152
≥ 9	5	7.9	10.9	45	405
	300	300.0	300.0	1324	6970

The frequencies f_c are calculated by the formula

$$f_c = \frac{300 \lambda^x e^{-\lambda}}{x!}$$

for the assumed value of λ . If Molina's tables (Reference 1) are not available, $e^{-\lambda}$ can be found from a table of e^{-x} (or by logarithms — the common

TABLE 37. SHORT TABLE OF $e^{-\lambda}$

λ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Factor
0	1.0000	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	1
1	.3679	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	1
2	.1353	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	1
3	.4979	.4505	.4076	.3688	.3337	.3020	.2732	.2472	.2237	.2024	10^{-1}
4	.1832	.1657	.1500	.1357	.1228	.1111	.1005	.0910	.0823	.0745	10^{-1}
5	.6738	.6097	.5517	.4992	.4517	.4087	.3698	.3346	.3028	.2739	10^{-2}
6	.2479	.2243	.2029	.1836	.1662	.1503	.1360	.1231	.1114	.1008	10^{-2}
7	.912	.825	.747	.676	.611	.553	.500	.453	.410	.371	10^{-3}
8	.335	.304	.275	.249	.225	.203	.184	.167	.151	.136	10^{-3}
9	1.23	1.12	1.01	.91	.83	.75	.68	.61	.55	.50	10^{-4}

To find $e^{-4.4}$, look under row 4 and column 0.4, and multiply the entry by the factor in the last column for row 4, namely 10^{-1} . Then $e^{-4.4} = 0.01228$.

logarithm of e is 0.4342945) and the successive values of f_c found by a recursion formula. The formula is

$$(10.26) \quad \frac{f_c(x+1)}{f_c(x)} = \frac{\lambda^{x+1}}{(x+1)! \lambda^x} \frac{x!}{x+1} = \frac{\lambda}{x+1}$$

Then

$$f_c(0) = 300 e^{-\lambda}$$

$$f_c(1) = \lambda f_c(0)$$

$$f_c(2) = \frac{\lambda}{2} f_c(1)$$

$$f_c(3) = \frac{\lambda}{3} f_c(2)$$

and so on. In Table 36, columns 3 and 4, these are calculated for $\lambda = 50/12$, and for $\lambda = 4.413$ (the value estimated from the distribution itself). For use in this example and the exercises at the end of the chapter a short table of $e^{-\lambda}$ is given in Table 37.

To calculate $e^{-\lambda}$ for $\lambda = 4.413$, we find by linear interpolation in the table, between the values for 4.4 and 4.5, that $e^{-\lambda} = 0.0121$, so that $f_c(0) = 3.63$. By logarithms, $\log_{10} e^{-\lambda} = -4.413 \times 0.43429 = -1.9165 = \bar{2}.0835$, so that $e^{-\lambda} = 0.0121$, checking the result from the table.

10.9 Poisson Distribution for Random Events. The Poisson distribution arises in a number of problems in which events occur over an interval of time in a random way. For example, in a collection of atoms of a radioactive element there will be, *on the average*, a certain number N which will disintegrate in a time T . The probability that exactly x atoms will disintegrate in time T is

$$P(x, N) = N^x e^{-N} / x!$$

This may be proved on the assumption that the disintegrations are both individually and collectively random. This means that the splitting of one atom has no effect on the chances of disintegration for any other atom, and also that the rate of decay of the radioactive substance is sufficiently slow that the probability that an atom will disintegrate in a small time δt is independent of what may have happened during any preceding time.

The same problem arises at a telephone switchboard, where the random events are incoming calls. It may be assumed that the calls are independent of each other. The assumption of *collective* randomness is not so clear, because there are busy periods and slack periods (such as lunch intervals), but if the time T minutes is not too long we can suppose that the number of calls in one minute during this time is independent of what has happened in earlier minutes of the same period. On these assumptions the probability

that in a time t minutes there will be n calls is

$$P(n, kt) = (kt)^n e^{-kt} / n!$$

where $k = N/T$, the average number of calls per minute during the whole period.

The same theory applies to the rate at which "clicks" are heard in a Geiger counter, when placed near a constant source of radioactivity. The data in Table 38 were obtained for the number of clicks x registered in 10-second periods in a physics laboratory, only the general "background" radiation being used. The mean number per 10-second period was 7.952, and the calculated values of the expected number of times that 0, 1, 2, ... clicks would be registered in 10 seconds are given in column 3. The general agreement of the theoretical Poisson distribution with the observed distribution is clear, and, as we shall see later, this agreement is quite as good as we could expect even if the assumption of a Poisson distribution for the population of 10-second intervals is really true.

TABLE 38. GEIGER COUNTER READINGS (BACKGROUND) IN 10-SECOND INTERVALS

x	f_0	f_c	x	f_0	f_c
0	0	0.2	11	35	35.4
1	1	1.4	12	26	23.5
2	5	5.6	13	16	14.4
3	13	14.8	14	5	8.2
4	23	29.3	15	1	4.3
5	54	46.6	16	1	3.9
6	66	61.8	17	1	
7	72	70.2	18	2	
8	64	69.8	19	0	
9	67	61.6	20	1	
10	47	49.0		500	500.0

Exercises

1. In the identity

$$\left(\frac{3}{5} + \frac{2}{5}\right)^6 = \sum_{x=0}^6 C(6, x) \left(\frac{2}{5}\right)^x \left(\frac{3}{5}\right)^{6-x}$$

find the sum of the terms for which x has the values 1, 2, 3. *Ans.* 12,096/15,625.

2. If ten coins are tossed, what is the probability that there are (a) exactly 3 heads (b) not more than 3 heads? Assume that the probability of head with each coin is $\frac{1}{2}$.

Ans. (a) 120/1024; (b) 176/1024.

3. Assume that the probability that a bomb dropped from an airplane will strike a certain target is $\frac{1}{5}$. If six bombs are dropped find the probability that (a) exactly two will strike the target (b) at least two will strike the target. *Ans.* (a) 768/3125; (b) 1077/3125.

4. Find the term independent of x in the binomial expansion of $(x - 1/x^2)^{3n}$.

Ans. $(-1)^nC(3n, n)$.

Hint. Write down the $(p + 1)$ th term (the term in $1/x^{2p}$) and choose p so that the exponent of x in the whole term is 0.

5. Prove that the greatest value of $C(2n, x)$ is when $x = n$.

6. Show that the number of permutations which can be formed from $2n$ letters, all of these letters being a's or b's, is greatest when the number of a's is equal to the number of b's.

Hint. If there are x a's the number of permutations is $C(2n, x)$. Use exercise 5.

7. An anti-aircraft battery had 3 out of 5 successes in shooting down "flying bombs" that came within range. What is the chance that if 8 bombs came within range not more than 2 got through?

Ans. 0.316.

8. (a) Find the values of $C(15, x)$ for $x = 0$ to $x = 15$, by writing out 5 additional rows in Fig. 34 (Pascal's Triangle).

(b) Evaluate the terms corresponding to $x = 5$ and $x = 6$ in $C(15, x) \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{15-x}$.

(c) Show that the expression in (b) is the probability of throwing x fives or sixes with 15 dice.

9. If the probability of x is as given in Exercise 8, what is the expected value of x and its standard deviation? What is the skewness of the distribution?

10. Assume that 0.04 is the theoretical rate of mortality in a certain age group. Suppose an insurance company is carrying $s = 1000$ such cases. What is the standard deviation of the death rate (i.e., of x/s , where x is the number of deaths)? What would it be if $s = 10,000$?

Hint. If the expectation of x is $s\theta$, the expectation of x/s is θ . If the variance of x is $s\theta(1 - \theta)$, the variance of x/s is $[s\theta(1 - \theta)]/s^2 = \theta(1 - \theta)/s$. See (11.1).

11. Four thumbtacks were tossed on a table 100 times. The number x that fell point up in each of the 100 trials was noted. The results were:

x	0	1	2	3	4
f_o	4	30	36	25	5

(a) Estimate the probability θ that a thumbtack will fall point up.

(b) Fit a binomial distribution, with the estimated θ , to the observed distribution, by calculating the expected frequencies in 100 trials, f_e , corresponding to the observed f_o .

12. A factory turns out articles of a standardized type at the rate of 1000 per day. Experience shows that on the average 0.2% of each day's production is defective. Show that it is rather unusual for any day's production to include more than four defective articles.

13. Assume that the chance of an individual coal miner being killed in a mine accident during a year is $\frac{1}{1400}$. Use the Poisson law to calculate the probability that in a mine employing 350 miners there will be at least one fatal accident in a year. *Ans.* 0.22.

14. A retailer with limited storage space finds that, on the average, he is able to sell 2 boxes of parrot food per week. He replenishes his stock every Monday morning so as to start each week with four boxes on hand.

(a) What is the probability that he sells his entire stock in a week?

(b) What is the probability that he is unable to fill at least one order?

(c) With how many boxes should he start the week to ensure that the probability of being able to fill all orders shall be at least 0.99? *Ans.* (a) 0.143; (b) 0.053; (c) 6.

15. The probability that a man aged 35 will die before reaching the age of 40 may be taken as 0.018. Out of a group of 40 men, now aged 35, what is the approximate probability

that x will die within the next 5 years? Draw up a table of the probabilities for different values of x .

16. (Wallis, quoted by Wilks). The following table shows the distribution of the numbers of vacancies occurring per year in the U. S. Supreme Court during the years 1837 to 1932.

<i>Vacancies per Year</i>	<i>Frequency</i>
0	59
1	27
2	9
3	1

Fit a Poisson distribution to this observed distribution.

17. (Bortkiewicz). A classical example of the Poisson distribution of rare events is that of the deaths of Prussian cavalry soldiers from the kicks of horses during the twenty years 1875-1894. The frequency distribution of the number of such deaths in 10 army corps, per corps per annum, was

<i>Deaths</i>	<i>Frequency</i>
0	109
1	65
2	22
3	3
4	1

Show that the mean number of deaths per corps per annum was 0.61. Fit a Poisson distribution and calculate the theoretical frequencies. (The agreement turns out to be very good.)

References

1. E. C. Molina, *Poisson's Exponential Binomial Limit* (D. Van Nostrand Co., Inc., 1942).
2. T. C. Fry, *Probability and Its Engineering Uses* (D. Van Nostrand Co., Inc., 1928). This book contains a good discussion of the Binomial and Poisson laws, with practical applications, and also includes tables of the Poisson function.

CHAPTER XI .

SIGNIFICANCE TESTS FOR BINOMIAL POPULATIONS

11.1 Approximation of the Binomial Distribution by a Normal Distribution.

For many problems connected with binomial distributions the exact solutions are cumbersome and tedious to calculate. Good approximations, however, may often be obtained by making use of the fact that, for large values of s and for values of θ not too near 0 or 1, the binomial distribution may be replaced for practical purposes by a normal distribution.

Consider for example, the following problem: A coin is tossed 100 times. What is the probability that the number of heads will lie between 40 and 60 inclusive?

The probability of exactly x heads in 100 tosses, assuming that the probability of head on a single toss is $\frac{1}{2}$, is $C(100, x)(\frac{1}{2})^{100}$. The required probability is therefore

$$\sum_{x=40}^{60} \frac{100!(\frac{1}{2})^{100}}{x!(100-x)!}$$

and so is equal to the sum of 21 terms. The exact calculation and addition of all these terms is laborious.

If the probability θ of success in a single trial is $\frac{1}{2}$, the mean of x in s trials is $s/2$ and the standard deviation is $(s/4)^{1/2}$. As s increases, therefore, the mean

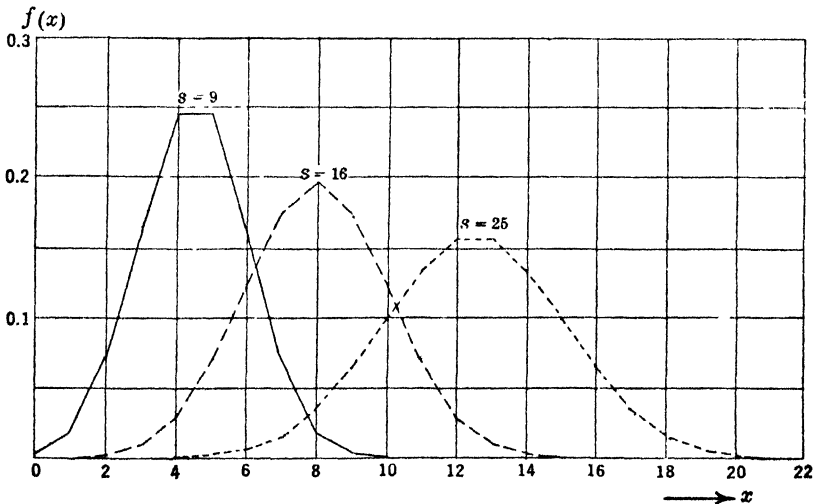


FIG. 39. BINOMIAL DISTRIBUTION ($\theta = 1/2$)

of the distribution increases and the dispersion also increases. The distribution moves to the right along the x axis and grows flatter, since the total area remains equal to unity. This is illustrated in Fig. 39 drawn for $s = 9, 16$ and 25. If, however, we standardize the variable by changing to

$$z = (x - s/2)/(s/4)^{1/2} = 2xs^{-1/2} - s^{1/2}$$

and keep the area constant by multiplying the ordinates of the distribution by $(s/4)^{1/2}$, we avoid this change in position and flattening out, and the distribution approaches nearer and nearer, as s increases, to the familiar shape of the normal curve. (See Fig. 40). For clearness, the figures are drawn as frequency polygons instead of histograms.

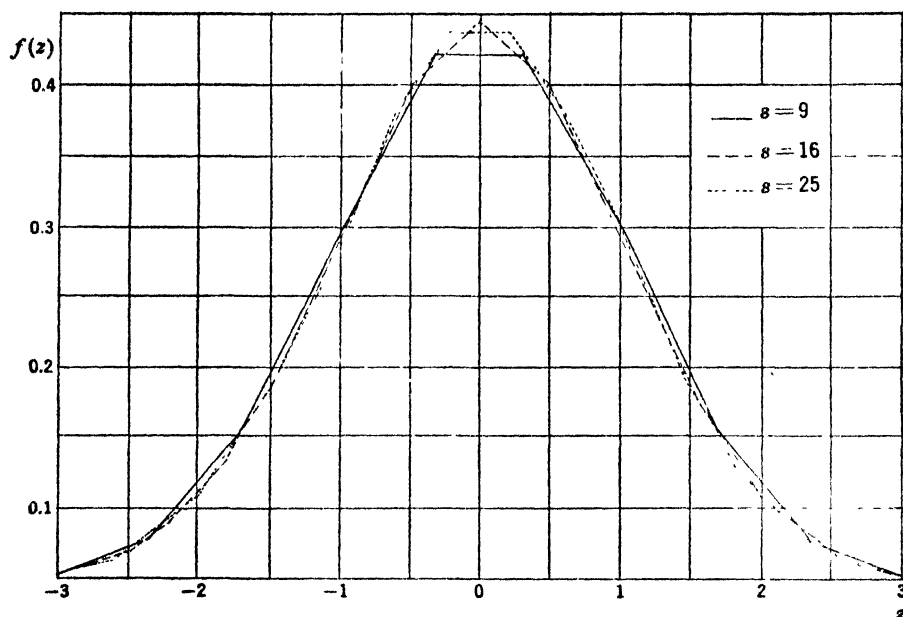


FIG. 40. BINOMIAL DISTRIBUTION (STANDARDIZED)

The rigorous proof that the limiting form of the binomial law is the normal law requires quite advanced mathematics. A proof which, while not rigorous, is illuminating and can be followed by the student who knows a little calculus may be found in Part Two, Chapter II. Here we shall simply accept the result that for large values of s the histogram of the binomial distribution is approximated by a normal curve, and that the approximation is much better when θ is near $\frac{1}{2}$ than when it is near 0 or 1.

Each term of the binomial distribution is represented in the histogram by a rectangle of unit base and height equal to $f(x)$. The sum of the terms from

$x = a$ to $x = b$ inclusive is therefore the combined area of the histogram from $x = a - \frac{1}{2}$ to $x = b + \frac{1}{2}$, since the base of the rectangle corresponding to a stretches from $a - \frac{1}{2}$ to $a + \frac{1}{2}$ and the base of the rectangle corresponding to b stretches from $b - \frac{1}{2}$ to $b + \frac{1}{2}$. This area is shaded in Fig. 41. If now the histogram is approximated by a normal curve with the same mean and standard deviation, and of course the same area, the sum of the binomial terms from $x = a$ to $x = b$ may be approximated by the area under the normal curve from $x = a - \frac{1}{2}$ to $x = b + \frac{1}{2}$. The mean of the normal curve will be $\mu = s\theta$, and the variance will be $\sigma^2 = s\theta(1 - \theta)$. The area from $a - \frac{1}{2}$ to $b + \frac{1}{2}$ will be the corresponding area under the standard normal curve from $z_1 = (a - \frac{1}{2} - s\theta)/\sigma$ to $z_2 = (b + \frac{1}{2} - s\theta)/\sigma$, and this is

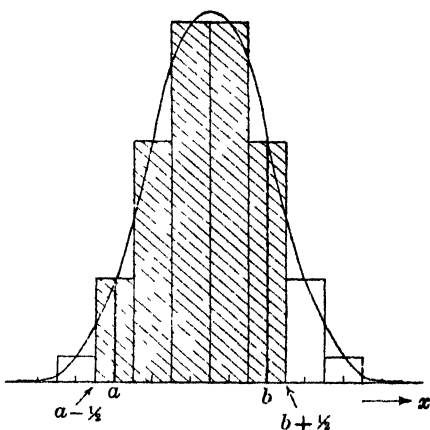


FIG. 41

$$\int_{z_1}^{z_2} \phi(z) dz = \Phi(z_2) - \Phi(z_1)$$

For the problem near the beginning of this section, concerning the probability of a number of heads between 40 and 60 inclusive in 100 tosses of a coin, we have $s\theta = 50$, and $\sigma = (100/4)^{1/2} = 5$, so that

$$z_1 = (40 - 0.5 - 50)/5 = -2.1$$

$$z_2 = (60 + 0.5 - 50)/5 = 2.1$$

The approximate probability is

$$\int_{-2.1}^{2.1} \phi(z) dz = 2 \int_0^{2.1} \phi(z) dz = 0.964$$

Example 1. Find the sum of the terms of $\left(\frac{3}{5} + \frac{2}{5}\right)^6$ for $x = 1, 2, 3, 4$ and compare with the normal approximation

By the binomial theorem,

$$\left(\frac{3}{5} + \frac{2}{5}\right)^6 = \sum_{x=0}^6 C(6, x) \left(\frac{2}{5}\right)^x \left(\frac{3}{5}\right)^{6-x}$$

so that the terms for $x = 1, 2, 3, 4$ are

$$\begin{aligned} 6 \binom{2}{5} \left(\frac{3}{5}\right)^5 &= 2916/15625 \\ 15 \binom{2}{5}^2 \left(\frac{3}{5}\right)^4 &= 4860/15625 \\ 20 \binom{2}{5}^3 \left(\frac{3}{5}\right)^3 &= 4320/15625 \\ 15 \binom{2}{5}^4 \left(\frac{3}{5}\right)^2 &= 2160/15625 \end{aligned}$$

the sum being $14256/15625 = 0.9124$.

For the normal approximation, $\mu = 6 \times \frac{2}{5} = 2.4$, $\sigma^2 = 6 \times \frac{2}{5} \times \frac{3}{5} = 1.44$, so that $z_1 = \frac{0.5 - 2.4}{1.2} = -1.58$ and $z_2 = \frac{4.5 - 2.4}{1.2} = 1.75$. The area from z_1 to z_2 is $0.95994 - 0.05705 = 0.9029$. The error is about 1%, which shows that the normal approximation is not bad even for as small a value of s as 6.

11.2 Significance of an Observed Proportion. Suppose that a population is divided into two categories, those which have and those which do not have a certain characteristic. For example, the population may consist of all registered births in 1952 in the state of Wisconsin, and the characteristic may be that of being male. Or the population may be tosses of a coin and the characteristic "heads." If in the whole population there is a proportion θ of individuals having this characteristic, the probability is θ that an individual picked at random will have the characteristic. The probability that in a sample of s individuals picked at random there will be x with the characteristic in question, is given by the binomial function

$$f(x) = C(s, x)\theta^x(1 - \theta)^{s-x}$$

Now the value of θ is usually unknown, but it can be estimated from the sample, and the larger the sample, the better will be the estimate. We will suppose that p is the *proportion* of individuals in the sample having the characteristic so that $p = x/s$. Obviously, if the sample s is so large that it includes the whole population, p will coincide with θ , but for small samples p will vary from one sample to another.

The expectation of x is, as we have seen, given by $\mu = s\theta$, so that, since s is a fixed number (the sample size), the expectation of p is $\frac{\mu}{s} = \theta$. The variance of p is

$$\begin{aligned} & \sum_{x=0}^s \binom{x}{s}^2 f(x) - \theta^2 \\ &= \frac{1}{s^2} \left[\sum_{x=0}^s x^2 f(x) - \mu^2 \right] = \frac{1}{s^2} \text{Var } (x) \\ (11.1) \quad &= \frac{1}{s^2} s\theta(1 - \theta) = \theta(1 - \theta)/s \end{aligned}$$

since, by (10.16), $\text{Var}(x) = s\theta(1 - \theta)$. The variance of p naturally diminishes as the sample size increases.

If now we *assume*, on grounds of general knowledge or experience or in virtue of some hypothesis which we wish to test, that θ has a certain value, we can estimate the probability that a sample of size s , taken at random from the population, will have a value of p differing from θ by any given amount. If the sample is large enough, the normal approximation to the binomial will enable us to estimate this probability quite easily. Thus, suppose we want to test the hypothesis that the probability of a male birth in Alberta is 0.5 exactly. We examine the registration statistics of the last 200 births (excluding stillbirths) in the province and find that there were, say, 110 males. The question arises whether this discrepancy from the expected value of 100 is large enough to make us doubt the theory that $\theta = 0.5$.

To answer this question we calculate the probability that, if θ is really 0.5, a random sample of 200 births would give a value of p at least as different from 0.5 as the value of 0.55 which we actually found, or in other words, that the number x of males in the 200 births would be at least as great as 110 or at least as small as 90. Now the probability that $x = 110$ or more is approximately the probability that

$$z > \frac{109.5 - 100}{(200 \cdot \frac{1}{2} \cdot \frac{1}{2})^{1/2}} = \frac{9.5\sqrt{2}}{10} = 1.3435$$

This probability is 0.0896, and there is an equal probability that $x = 90$ or less. The probability of a deviation from expectation at least as great as that found in the sample, on the hypothesis that the true value of θ is 0.5, is therefore about 0.18 or a little more than $\frac{1}{5}$, which is hardly small enough for us to feel safe in rejecting the hypothesis that θ is really 0.5. We say in such a case that the deviation from expectation is *non-significant*. Where we draw the line is a matter of convention.

Statisticians have more or less agreed to regard a deviation from expectation as *significant* if the probability of obtaining by pure chance a deviation at least as great as this is less than 0.05, and to regard it as *highly significant* if the probability is less than 0.01. To some extent, of course, the limits of significance will depend on the seriousness of making a mistake. If it is going to be a costly matter to reject wrongly some hypothesis about θ , we shall want to be very sure that we are right, and perhaps we shall want the probability of a chance deviation as great as that observed to be less than 0.005. However, in most of the problems in this book we shall accept the conventional significance levels of 0.05 and 0.01. If the probability calculated is less than 0.05, we shall say that the hypothesis regarding θ is rejected *at the 0.05 level of significance*. If the probability is less than 0.01, we shall say that the hypothesis is rejected *at the 0.01 level of significance*. If the prob-

ability is greater than 0.05, we shall say that the hypothesis is not rejected, at the 0.05 level. This is the situation in our example. On the basis of the sample, we could not reasonably reject the hypothesis that male and female births have an equal probability. Of course, new and larger samples might very well cause us to revise our opinion.

11.3 Tests of Hypotheses. *One-tailed and Two-Tailed Tests.* In the example discussed in §11.2, a hypothesis was made about the parameter θ , namely, that $\theta = 0.5$. This is called a *simple hypothesis*, because it fixes the binomial distribution completely (for a given sample size s). A hypothesis which does not completely specify the distribution is said to be *composite*. The proportion p of individuals in a sample having the characteristic under discussion (of which the probability in the population was θ) was used to test the assumed value of θ . The method adopted was to obtain the probability distribution for p and to calculate (at least approximately) the probability of getting, in a random sample of size s , a proportion *at least as different* from θ as the observed p itself. This involved finding the areas under the standard normal curve, both for z greater than a certain value z_1 and for z less than $-z_1$, where z_1 was calculated from p . Since these areas correspond to the upper and lower "tails" of the normal curve, the test is often referred to as a *two-tailed test*. We always use a two-tailed test if we are interested in the *amount* of the deviation from expectation, without caring very much in what direction it goes, but situations do arise sometimes where we have good reason to believe, *before we perform the experiment or obtain the observations*, that the deviation, if it occurs, will be in one direction only. If so, we are justified in using a *one-tailed test*.

Example 2. The same test is given to 100 subjects twice, and 60 of these get a better score on the second attempt than on the first. Is this fact significant?

The hypothesis tested is that the repetition makes no difference to the scores. Let us suppose we have decided in advance that the only possible effect of repetition, if there is any effect at all, must be an *improvement* in the scores. The question then arises as to the *significance* of the amount by which the observed p *exceeds* 0.5. The equivalent normal value is $(59.5 - 50)/(25)^{1/2} = 1.9$, and the area under the normal curve for $z > 1.9$ is 0.029. That is to say, the probability of 60 or more out of 100 getting an improved score by chance, if actually repetition makes no difference, is about 0.029, and so the result is judged *significant* by the usual criterion. If we used a two-tailed test, the probability that the number *getting* a better score is *either* 60 or more *or* 40 or less would be 0.058, and this would usually be judged non-significant.

A hypothesis that is tested for possible rejection, such as the hypothesis that $\theta = 0.5$ in Example 2, is called a *null hypothesis*. This is tested against an *alternative hypothesis*, which, in this example, is that $\theta > 0.5$ (a one-sided alternative). On the basis of the observed results, the null hypothesis is rejected at the 5% level of significance, and therefore the alternative hypothesis is accepted at the same level. Notice that the null hypothesis is not

disproved. All we can say is that *either* the null hypothesis is untrue *or* that a very unusual event has occurred, namely, an event with a probability of less than 1 in 30. We shall usually prefer to make the former statement, but if we do so we stand a chance, even though a small one, of being wrong.

11.4 Confidence Limits for the Binomial Parameter. Instead of using the sample to test an assumed value of θ , we can regard it as giving an *estimate* of an otherwise unknown θ . Since the expectation of p is θ itself, p provides an *unbiased estimate*. We know, however, that p will vary from one sample to another, even though the sample size remains constant, and therefore our estimate is subject to sampling error. In fact, the sampling variance of p is $\theta(1 - \theta)/s$. For any observed value of p we can set limits which will be wide enough to include the true value of θ with any required degree of confidence, say, 0.95. These limits are called the *95% confidence limits* for θ , for a reason that will now be explained.

The true value θ , although unknown, is not a random variable, so that we should not speak of the probability that θ lies between the confidence limits. Rather it is p which is the random variable, and there is a probability that the *confidence interval* (between the upper and lower confidence limits), which is a function of p , will *include* the true value θ . If we take many samples and calculate many 95% confidence intervals, then about 95% of them will include the true value, so that in saying of one confidence interval that it does include the true value we stand only a 5% chance of being wrong, and are therefore reasonably confident of being right.

We now consider how to calculate these limits, and once again we appeal to the normal approximation to the binomial distribution. We know that $z = (sp - s\theta)/[s\theta(1 - \theta)]^{1/2}$ is approximately $N(0, 1)$, and that the two symmetrical values of z which between them include 95% of the whole area of the normal distribution are ± 1.96 . (The area of the tail beyond $z = 1.96$ is 0.025, and there is an equal area in the other tail below -1.96 .) There is therefore a probability of 0.95 that, if θ is the true value of the parameter,

$$(11.2) \quad -1.96 \leq \frac{sp - s\theta}{[s\theta(1 - \theta)]^{1/2}} \leq 1.96$$

This inequality can be written as

$$(sp - s\theta)^2 \leq (1.96)^2 s\theta(1 - \theta)$$

or

$$(11.3) \quad s(p - \theta)^2 \leq 3.84 \theta(1 - \theta)$$

Collecting together the terms in θ and θ^2 , we obtain

$$(11.4) \quad \theta^2(s + 3.84) - \theta(2ps + 3.84) + sp^2 \leq 0$$

Now the quadratic expression $a\theta^2 + b\theta + c$ is negative, for $a > 0$, when θ lies between the limits $(-b \pm \sqrt{b^2 - 4ac})/2a$, which are the roots of the equation obtained by replacing the inequality by an equality. The upper and lower confidence limits are therefore given by the two roots of the quadratic (11.4) with the left-hand side equal to 0. One root corresponds to the right-hand inequality in (11.2) and the other to the left-hand inequality. A general expression can be written for these roots, but the procedure is best illustrated by a numerical example.

Example 3. In a sample of 200 persons interviewed in a public opinion poll, 124 said "yes" to a certain question and the remainder said "no" (For simplicity we are ignoring those persons who were undecided.) What are the 95% confidence limits for the proportion of persons in the population sampled (supposed very large) who would answer "yes"?

Here $p = 0.62$, $s = 200$, and the quadratic equation for θ is $203.84\theta^2 - 251.84\theta + 76.88 = 0$. The roots are $\theta = 0.551$ and 0.684 , which provide the required confidence limits.

When s is fairly large, as in this example, it is unnecessary to correct the observed x by adding or subtracting $\frac{1}{2}$ in forming the normal approximation. If greater accuracy is desired, however, we must replace the inequalities in (11.2) by the following pair:

$$(11.5) \quad \begin{cases} \frac{sp - \frac{1}{2} - s\theta}{[s\theta(1-\theta)]^{1/2}} \leq 1.96 \\ \frac{sp + \frac{1}{2} - s\theta}{[s\theta(1-\theta)]^{1/2}} \geq -1.96 \end{cases}$$

The first gives $\theta \geq 0.549$ and the second $\theta \leq 0.687$, so that the confidence limits are only slightly affected.

If we want 90% confidence limits, the number 1.96 in (11.2) or (11.5) must be replaced by 1.645, and if we want 99% confidence limits by 2.576. The values for other confidence limits can be obtained from tables of the normal law, but these are the ones usually adopted.

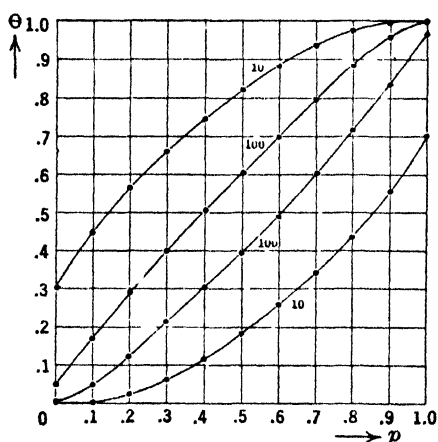


FIG. 42

11.5 Confidence Interval Charts.

A convenient set of charts for reading off the two values of θ corresponding to given values of p and s has been devised by Clopper and Pearson (Reference 1). For each value of s two curves connecting p and θ are drawn, as in Fig. 42. The ordinates at any given p are the corresponding upper and lower confidence limits. For the larger values of s the curves are drawn from equation (11.2), but for small values of s the normal approximation is not adequate and the binomial function itself must be used. Recently published tables of the binomial

distribution (see Reference 2) make it easier to obtain the values of θ

for which

$$\sum_{x=sp}^s C(s, x) \theta^x (1 - \theta)^{s-x} = 0.025$$

and for which

$$\sum_{x=0}^{sp} C(s, x) \theta^x (1 - \theta)^{s-x} = 0.025,$$

these values being respectively the lower and upper confidence limits corresponding to the observed p . Thus if $s = 10$ and $p = 0.5$, we find from the tables that the limits are 0.187 and 0.813. The values given by (11.5) are 0.201 and 0.798, which differ appreciably from the true values.

Clopper and Pearson's chart for 95% confidence limits is reproduced as Chart I in the Appendix. Notice that the smaller the sample size, the wider the confidence limits. The 90% confidence limits are narrower than the 95% limits, because the more precise the statement about θ , the more chance there is of being wrong in making it.

11.6 Mean and Variance of a Linear Combination of Independent Variates.

In the next section we shall consider the significance of a difference in the observed proportions between two independent samples, and we shall need Theorem 1, which follows. A straightforward, but rather long, elementary proof can be given, but as the result is the important thing here, we refer the reader to a more sophisticated proof in Part Two, §4.11.

Theorem 1. *If x_1 and x_2 are independent variates with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , and if $L = c_1x_1 + c_2x_2$ is a linear combination of x_1 and x_2 , then the mean of L is $\mu_L = c_1\mu_1 + c_2\mu_2$ and the variance of L is*

$$\sigma_L^2 = c_1^2 \sigma_1^2 + c_2^2 \sigma_2^2.$$

For example suppose $c_1 = 1$ and $c_2 = -1$, then $L = x_1 - x_2$ and we have the result that

$$(11.6) \quad \sigma_L^2 = \sigma_1^2 + \sigma_2^2$$

11.7. Significance of a Difference Between Two Sample Proportions.

Suppose we have two independent random samples which may be of different sizes s_1 and s_2 , and suppose in each we observe the number of individuals, say x_1 and x_2 respectively, possessing a certain characteristic. Then the sample proportions are $p_1 = x_1/s_1$ and $p_2 = x_2/s_2$, and these will, in general, differ. We can test the null hypothesis that in fact both samples came from the same population in which the true proportion is θ .

On the null hypothesis, the expectations of p_1 and p_2 are both equal to θ , so that the expectation of $p_1 - p_2$ is zero. The variance of p_1 is $\theta(1 - \theta)/s_1$ and that of p_2 is $\theta(1 - \theta)/s_2$; since p_1 and p_2 are independent variates, the

variance of $p_1 - p_2$ is the *sum* of the respective variances, by the theorem in §11.6. Therefore, for large values of s_1 and s_2 , we can treat

$$z = (p_1 - p_2) / \left[\theta(1 - \theta) \left(\frac{1}{s_1} + \frac{1}{s_2} \right) \right]^{1/2}$$

as a standard normal variate, and estimate the probability of a value of z at least as great as that observed, for an assumed value of θ .

Example 4. In a referendum submitted to the student body at a university, 850 men and 566 women voted. 530 of the men and 304 of the women voted "yes." Does this indicate a significant difference of opinion on the matter, at the 1% level, between men and women students?

The null hypothesis is that the proportion θ who would vote "yes" in a very large student population is the same for men and women. We do not know the value of θ and can only estimate it from the combined results. The over-all proportion of students voting "yes" is $834/1416 = 0.589$, and this may be taken as an estimate of θ . Also, we know that $p_1 = 530/850 = 0.6235$, $p_2 = 304/566 = 0.5371$, so that

$$z = 0.0864 / \left[0.589 \times 0.411 \left(\frac{1}{850} + \frac{1}{566} \right) \right]^{1/2} = 3.24$$

Since any value of z greater than 2.576 indicates significance at the 1% level, the question is answered in the affirmative; in other words, the null hypothesis is rejected at this level.

11.8 Confidence Limits for the Difference in Proportions. If we do not wish to make the null hypothesis that the two samples are from the same population, we can assume that they come from two different populations with probabilities θ_1 and θ_2 , and obtain confidence limits for the difference $\theta_1 - \theta_2$. The variable

$$(11.7) \quad z = [p_1 - p_2 - (\theta_1 - \theta_2)] / [\theta_1(1 - \theta_1)/s_1 + \theta_2(1 - \theta_2)/s_2]^{1/2}$$

is approximately normally distributed, for large s_1 and s_2 , so that we can obtain 95% confidence limits by putting $z = \pm 1.96$. Unfortunately, θ_1 and θ_2 are unknown, but we can approximate them, in the denominator* of (11.7), by putting p_1 for θ_1 and p_2 for θ_2 . We have then the equation

$$(11.8) \quad p_1 - p_2 - (\theta_1 - \theta_2) = \pm 1.96 [p_1(1 - p_1)/s_1 + p_2(1 - p_2)/s_2]^{1/2}$$

for determining the two values of $\theta_1 - \theta_2$ which are the lower and upper confidence limits. Thus, in Example 4, the equation is

$$\begin{aligned} \theta_1 - \theta_2 &= 0.0864 \pm 1.96 [0.000439 + 0.000276]^{1/2} \\ &= 0.0864 \pm 0.0524 \\ &= 0.034 \text{ or } 0.139 \end{aligned}$$

The 95% confidence limits are therefore 0.034 and 0.139, and since these

* Since we want the difference between $\theta_1 - \theta_2$ and $p_1 - p_2$, we obviously cannot make the approximation in the numerator as well. The denominator does not contain this difference.

limits do not include zero, we can say that there is a significant difference at the 5% level between the opinions of men and women students on the subject of the referendum. The 99% limits are 0.018 and 0.155, and still do not include 0.

11.9 Binomial Probability Paper. A special graph paper, designed by Mosteller and Tukey (see Reference 3) enables problems on the significance of proportions to be solved approximately with great ease. A specimen of the paper is shown in Fig. 43. The scales are "square root" scales, that is

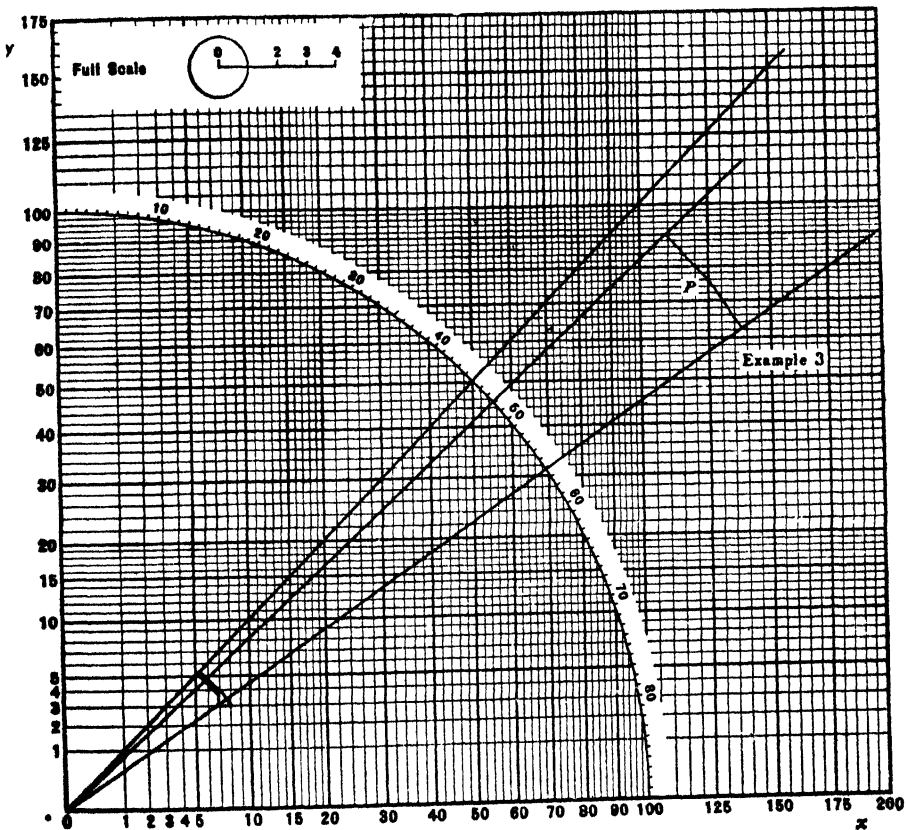


FIG. 43. BINOMIAL PROBABILITY GRAPH PAPER

to say, the distance of points marked x and y from the origin is proportional to $x^{1/2}$ and to $y^{1/2}$ respectively. A quarter-circle is drawn through the points marked 100 on the axis, and on this circle $x + y = 100$. A straight line through the origin passes through points for which y/x is constant and is called a *split*. Thus a 50-50 split is a line passing through the point on the quarter-circle of coordinates 50, 50.

Suppose in a sample of 10 we find 7 with a certain characteristic, which we will call "success." Since there are 3 in the sample which are "failures," we say that the *paired count* for the sample is (7, 3). This is plotted on the graph paper as a right-angled triangle, with the right angle at (7, 3) and sides each one unit long, parallel to the axes (see Fig. 43). When one of the coordinates is larger than about 100, the one unit length is scarcely more than the width of a pencil line, and the triangle becomes a short segment. If both coordinates are large, the triangle reduces to a point.

In order to test whether the observed proportion of successes ($\frac{7}{10}$) is significantly different from a hypothetical θ of $\frac{1}{2}$, we measure the perpendicular distance from the plotted triangle to the 50-50 split. When the numbers are small, there are two distances, called the *short* and the *long distance*, measured from the two acute angles of the triangle. These distances are interpreted by reference to the scale at the top of the paper (marked Full Scale). It may be noted that the distance 0 to 1 on this scale is almost exactly 5 mm. The long and short distances give each a significance level, and the observed result is significant at some level between. When both coordinates are large there is only one significance level. The scale is a normal probability one, giving values of z corresponding to the observed deviation. Thus, 2 units on this scale correspond very nearly to the 5% level of significance (for a two-tailed test). In the illustration above, the long and short distances are 1.0 and 1.6, so that the result is not significant at the 5% level.

In Example 3 of §11.4, the paired count is (124, 76). If this is plotted (as P in Fig. 43), and two splits are drawn at a perpendicular distance of 2 units from the triangle (practically a point), the coordinates of these splits provide upper and lower 95% confidence limits for θ . In the diagram these splits are (55, 45) and (69, 31) so that the upper and lower confidence limits are 0.55 and 0.69, agreeing quite well with the results calculated earlier.

This type of graph paper has many other uses which are explained and illustrated in the paper quoted as Reference 3.

11.10 Sampling from a Finite Population. In all our illustrations in this chapter, we have supposed that the parent population which is sampled may be taken as infinite, or practically so. However, examples sometimes occur of sampling from a finite population, of a size comparable with the sample, and then the formulas for testing significance must be modified.

It is proved in Part Two that if a sample of size s is drawn from a binomial population of size n , the variance of x (the number of "successes" in the sample) is equal to what it would be if the population were infinite, multiplied by the factor $(n - s)/(n - 1)$. The distribution of x in this case is called "hypergeometric." The mean of x is the same as in the pure binomial distribution. The only change we have to make in our formulas is therefore to replace $s\theta(1 - \theta)$ by $s\theta(1 - \theta)(n - s)/(n - 1)$.

Example 5. A telephone poll was taken of University of Alberta students before a student election. Of 78 students contacted, 33 said they would vote for Mr. S. The population (of students with telephones) may be taken as 2200. Obtain 95% confidence limits for the proportion of voters in the population in favor of Mr. S.

Here $s = 78$, $n = 2200$, $p = 33/78 = 0.423$. Instead of equation (11.2) we now have

$$(11.9) \quad -1.96 \leq \frac{sp - s\theta}{[s\theta(1 - \theta)]^{1/2}} \left(\frac{n - 1}{n - s} \right)^{1/2} \leq 1.96$$

or

$$(11.10) \quad s(p - \theta)^2 \leq 3.84 \frac{n - s}{n - 1} \theta(1 - \theta)$$

$$\text{Then} \quad 78(0.423 - \theta)^2 \leq 3.84 \times \frac{2122}{2199} \theta(1 - \theta)$$

whence the limits of θ are given by

$$81.7\theta^2 - 69.7\theta + 14.0 = 0$$

The roots are 0.32 and 0.53, which are the required confidence limits.

Exercises

1. If 500 coins are tossed, what is the probability that (a) the number of heads will differ from 250 by less than 20 (b) at least 260 coins will show heads?

2. In a certain game, called "Twenty-six," played with ten dice, a player chooses some number, such as 4, and undertakes to throw 26 or more fours in thirteen throws with the ten dice. If he succeeds the bank pays him 4 to 1. How much *should* the bank pay to make the game fair?

Ans. 4.43 to 1.

Hint. Find the normal approximation to the probability of 26 or more fours, with 130 throws of a die (equivalent to 13 throws with 10 dice). If the player stakes \$1 and the bank \$M, and if the probability of success is p , then for a fair game $pM - (1 - p) \cdot 1 = 0$.

3. Out of a very large Grade 12 population of high school students, 40% fail a university entrance examination. In a certain class of 50 the number of failures was 14. Is this deviation from the general average large enough to warrant a presumption that the class is not a random sample from the Grade 12 population?

4. Two groups of mice, one of 50 and the other of 60, are comparable in respect of age, weight, general condition, etc., and have both been given injections of a virus. The first group, however, has also been given a certain drug. After three days, the number of deaths in the first group was 12 and in the second 19. Is the difference in mortality rates significant?

5. 400 eggs are taken at random from a large consignment and 50 are found to be bad. What are the 99% confidence limits for the percentage of bad eggs in the whole consignment?

6. For a certain year in Canada the deaths recorded for married men and single men respectively in the age group 25 to 44 were 3471 and 2307. If the whole population in this age group numbered 377,000 single men and 940,000 married men, could the single men be reasonably regarded as a random sample of the whole male population, as far as mortality rate is concerned?

7. In a large city 400 voters were chosen at random and asked whether they would vote for Candidate A at the next election; 280 said they would. What are the 99% confidence limits for the proportion of voters in the city who intend to vote for A?

8. A physician treats 20 patients suffering from a certain disease and 11 of them die. The mortality rate in this disease, based on thousands of reported cases, is 42%. Can the sample be regarded as exceptional?

9. (*Mavland*) In testing the efficacy of a drug said to prevent seasickness, 25 men who always developed symptoms of sickness when subjected to the motion of a rocking machine were given the drug. Again tested, 15 were now found to be immune to the motion. What are the 95% confidence limits for the proportion of men liable to seasickness who would be rendered immune by taking this drug?

10. In examining the pedigrees of a number of families the occurrence of a particular hereditary defect is observed in 14 out of 25 children. If a certain genetic mechanism is at work, the proportion of these children affected should be $3/4$. Are the data consistent with this mechanism?

11. (*Crespi*, quoted by *Wilks*) In a poll of 148 men and 152 women the question was asked, "Do you approve of the practice of tipping, by and large?" and 89 of the men and 116 of the women answered "yes". Construct 95% confidence limits for the difference between the proportion of "yeses" among the male population sampled and the proportion among the women sampled. (Assume that these populations are very large compared with the sample sizes.)

12. Random samples of 50 students each from the freshman class in Arts and Science and the freshman class in Engineering are given a mathematical aptitude test. The numbers reaching a pass standard are 35 and 41, respectively. Assuming that the whole Arts and Science class includes 248 students and the Engineering class 187, test at the 5% level the null hypothesis that the proportion of successes is the same in both classes.

References

1. C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits," *Biometrika*, **26**, 1934, pp. 404-413. These charts (and additional ones for 80% and 90% confidence limits) are reproduced in *Introduction to Statistical Analysis*, by W. J. Dixon and F. J. Massey (McGraw-Hill Book Co., Inc., 1950).

2. *Tables of the Binomial Probability Distribution* (National Bureau of Standards, Washington, 1949). These tables give the individual terms $C(n, r)p^r q^{n-r}$ for values of n from 2 to 49, r from 0 to $n-1$, and p at every 0.01 from 0.01 to 0.50. They also give the sums $\sum_{s=r}^n C(n, s)p^s q^{n-s}$. See also H. G. Romig, *50-100 Binomial Tables*, J. Wiley and Sons, Inc., New York, 1953.

3. F. Mosteller and J. W. Tukey, "The Uses and Usefulness of Binomial Probability Paper," *J. Amer. Stat. Assoc.*, **44**, 1949, pp. 174-212. The graph paper is obtainable from the Codex Book Co., Inc., Norwood, Mass.

4. Reference 18 in §0.4 contains a simple treatment of levels of significance, confidence intervals, etc.

CHAPTER XII

SIGNIFICANCE OF MEANS AND VARIANCES

12.1 Distribution of the Sample Mean. It is proved in Part Two that, if a great number of random samples of size N are picked from an infinitely large population, and if the arithmetic mean \bar{x} is computed for each sample, then \bar{x} has a probability distribution for which the moments can be calculated in terms of the moments of the parent population. If we denote by $\mu_{\bar{x}}$, $\sigma_{\bar{x}}^2$, $\gamma_{1,\bar{x}}$, etc., the mean, variance, skewness, etc., for the distribution of the mean, and if we let μ , σ^2 , γ_1 , etc., denote the corresponding quantities for the parent population, then we can show that

$$(12.1) \quad \mu_{\bar{x}} = \mu$$

$$(12.2) \quad \sigma_{\bar{x}}^2 = \sigma^2/N$$

$$(12.3) \quad \gamma_{1,\bar{x}} = \gamma_1/N^{1/2}$$

$$(12.4) \quad \gamma_{2,\bar{x}} = \gamma_2/N$$

The mean of all the means is therefore equal to the mean of the parent population, but the standard deviation and the skewness are equal to the corresponding population values divided by \sqrt{N} , while the kurtosis is the population value divided by N . It follows that if N is fairly large and if γ_1 and γ_2 are moderate, the skewness and kurtosis of the distribution of means will be near to zero, and it appears that the distribution of means will somewhat resemble a normal distribution. If it happens that the parent population is itself normally distributed, then $\gamma_1 = \gamma_2 = 0$, so that the skewness and kurtosis of the distribution of means are also zero, and in fact the distribution of means is then itself exactly normal with mean μ and variance σ^2/N .

The results expressed by equations (12.1) to (12.4) are of great importance in estimating the significance of differences in the means between two samples and in finding confidence limits for the mean of the parent population from the mean of a sample. For proofs, the reader may refer to Part Two, §§4.16 and 6.8. Proofs involving only elementary algebra can be given, but are likely to be rather long and tedious. We give, to illustrate the type, the proof of (12.1) for a discrete distribution and for a sample of only two individuals.

Let us suppose that in the population the variate x can take the distinct values x_1, x_2, \dots, x_k , with probabilities p_1, p_2, \dots, p_k . For the sample, two

items are picked at random, say x_α and x_β (these may, of course, be equal) and the sample mean is

$$(12.5) \quad \bar{x} = (x_\alpha + x_\beta)/2$$

Since the two items are independent, the probability of picking just these two is the product of the probabilities for the two separately, namely, $p_\alpha p_\beta$. The mean $\mu_{\bar{x}}$ of the values of \bar{x} for all possible samples of two is given by multiplying \bar{x} by the probability of the sample and summing over all possible values of α and β , from 1 to k . That is,

$$(12.6) \quad \begin{aligned} \mu_{\bar{x}} &= \sum_{\alpha, \beta} \bar{x} p_\alpha p_\beta \\ &= \frac{1}{2} \sum_{\alpha, \beta} x_\alpha p_\alpha p_\beta + \frac{1}{2} \sum_{\alpha, \beta} x_\beta p_\beta p_\alpha \end{aligned}$$

Now $\sum_\alpha p_\alpha = \sum_\beta p_\beta = 1$, since each of these is simply the sum of the probabilities for all possible values of x . Also, by definition of the population mean,

$$(12.7) \quad \mu = \sum_\alpha x_\alpha p_\alpha = \sum_\beta x_\beta p_\beta$$

the two sums being the same since α and β both have the same domain of values 1 to k . Therefore, from (12.6),

$$(12.8) \quad \begin{aligned} \mu_{\bar{x}} &= \frac{1}{2} \sum_\beta \mu p_\beta + \frac{1}{2} \sum_\alpha \mu p_\alpha \\ &= \frac{1}{2} \mu \left(\sum_\alpha p_\alpha + \sum_\beta p_\beta \right) = \mu \end{aligned}$$

which is the same as (12.1).

The proof for the variance is similar, using the definition

$$(12.9) \quad \sigma_{\bar{x}}^2 = \sum_{\alpha, \beta} \bar{x}^2 p_\alpha p_\beta - \mu_{\bar{x}}^2$$

and

$$(12.10) \quad \sigma^2 = \sum_\alpha x_\alpha^2 p_\alpha - \mu^2 = \sum_\beta x_\beta^2 p_\beta - \mu^2$$

but we will not give it in detail. It turns out that $\sigma_{\bar{x}}^2 = \frac{1}{2}\sigma^2$, which is the same as (12.2) for the special case $N = 2$. The extension of these proofs to three, four, or more items in the sample should be obvious.

12.2 An Illustration of the Distribution of Means. An experiment in sampling was carried out in class as follows: A "population" was constructed by writing numbers from 0 to 24 on 1000 metal-edged cardboard tags. The distribution of the numbers was as given in Table 38, and it is clear that this distribution is markedly skew.

TABLE 38. PARENT POPULATION IN SAMPLING EXPERIMENT

x	f	x	f
0	1	13	32
1	23	14	26
2	61	15	20
3	92	16	16
4	106	17	13
5	100	18	10
6	94	19	8
7	87	20	6
8	78	21	4
9	69	22	3
10	59	23	2
11	49	24	1
12	40		
			1000

The numbered discs were put into a goldfish bowl and well mixed. A sample of 10 discs was withdrawn, and the numbers were noted before the discs were replaced. The process was then repeated until a few hundred samples had been obtained, and a frequency distribution of the means was constructed. Over a period of several years, the data summarized in Table 39 were collected.

For the population of Table 38, the following values may be calculated:

$$(12.11) \quad \begin{cases} \mu = 7.601 \\ \sigma^2 = 19.57 \\ \gamma_1 = 0.896 \\ \gamma_2 = 0.508 \end{cases}$$

For the distribution of Table 39, we find (using Sheppard's corrections)

$$(12.12) \quad \begin{cases} \mu_{\bar{x}} = 7.64 \\ \sigma_{\bar{x}}^2 = 1.985 \\ \gamma_{1,\bar{x}} = 0.343 \\ \gamma_{2,\bar{x}} = 0.002 \end{cases}$$

These are actually the sample statistics, for the sample of 2100 means, but this number is so large that the estimates for the whole population of possible means will not differ appreciably from the sample statistics. The results in (12.12) may be compared with the expected values,

$$\mu = 7.60, \quad \sigma^2/10 = 1.957, \quad \gamma_1/\sqrt{10} = 0.283, \quad \text{and} \quad \gamma_2/10 = 0.051$$

and are seen to be of about the right size. At the present stage we cannot

TABLE 39. DISTRIBUTION OF MEANS OF 2100 SAMPLES OF 10 FROM POPULATION OF TABLE 38

<i>Class Limits</i>	<i>f</i>	<i>x_c</i>	<i>u</i>	<i>fu</i>	<i>fu²</i>
3.0- 3.9	1	3.45	-4	-4	16
4.0- 4.9	27	4.45	-3	-81	243
5.0- 5.9	210	5.45	-2	-420	840
6.0- 6.9	463	6.45	-1	-463	463
7.0- 7.9	559	7.45	0	0	0
8.0- 8.9	477	8.45	1	477	477
9.0- 9.9	234	9.45	2	468	936
10.0-10.9	99	10.45	3	297	891
11.0-11.9	23	11.45	4	92	368
12.0-12.9	6	12.45	5	30	150
13.0-13.9	1	13.45	6	6	36
	2100			402	4420

go further in judging the agreement. The distributions of Tables 38 and 39 are graphed in Fig. 44, and this shows clearly the reduction in standard deviation and skewness brought about by taking the mean of even so small

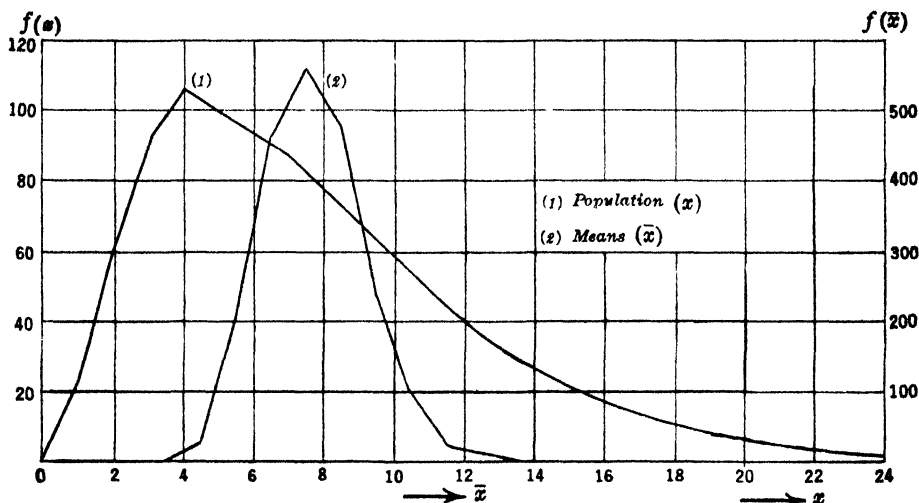


FIG. 44. DISTRIBUTION OF MEANS OF SAMPLES OF 10 FROM SKEW POPULATION

a sample as 10. The parent population is actually finite, but the correction for this is not important, since the population is 100 times as large as the sample. (See §12.6.)

12.3 Significance of Means. If we suppose that the mean \bar{x} of a sample of N is normally distributed with mean μ and variance σ^2/N , we can use the tables of the normal law to estimate the probability of a difference at least as great as that observed between \bar{x} and an assumed value of μ .

Example 1. If a "true" die is rolled 50 times, what is the probability that the average number of spots obtained will be at least 4?

The probability for x spots in a single throw is assumed to be $\frac{1}{6}$ for all values of x . This is implied in the statement that the die is true. The population consists of the infinitely many throws which are at least conceivable with the given die. For this population,

$$\mu = \sum_{x=1}^6 xf(x) = \frac{1}{6} \sum x = 3.5$$

and

$$\begin{aligned} \sigma^2 &= \sum_{x=1}^6 x^2 f(x) - \mu^2 = \frac{1}{6} \sum x^2 - (3.5)^2 \\ &= \frac{1}{36} (6 \cdot 7 \cdot 13) - (3.5)^2 = \frac{35}{12} \end{aligned}$$

(see Theorem 5, §4.2). Therefore if \bar{x} is the mean of x for a sample of 50,

$$\mu_{\bar{x}} = 3.5$$

$$\sigma_{\bar{x}}^2 = 35/(12 \cdot 50) = 7/120 = 0.05833$$

If the distribution of \bar{x} is normal, the standard normal variate will be

$$z = (\bar{x} - 3.5)/(0.05833)^{1/2} = (\bar{x} - 3.5)/0.2415$$

For $\bar{x} = 4$,

$$z = 0.5/0.2415 = 2.07$$

and the probability of a value of z at least as great as 2.07 is 0.0192. This is the probability required.

Note that from the wording of the question we are concerned only with the upper tail of the normal distribution. If we want the probability that the observed mean differs from the expected value 3.5 by at least 0.5 either way, the probability obtained above must be doubled.

12.4 Confidence Interval for Means. There are two parameters in the equation of the normal law, μ and σ , but if we suppose that σ is fixed, we can estimate confidence limits for μ from the mean \bar{x} of a sample. Since \bar{x} is normal with mean μ and standard deviation $\sigma/N^{1/2}$, we have only to put

$$(12.13) \quad z = N^{1/2}(\bar{x} - \mu)/\sigma = \pm 1.96$$

to obtain the 95% confidence limits for μ . This equation may be written

$$(12.14) \quad \mu = \bar{x} \pm 1.96 \sigma/N^{1/2}$$

and it is clear that the width of the confidence interval (between the upper and lower confidence limits) is constant for all values of \bar{x} , as illustrated in Fig. 45.

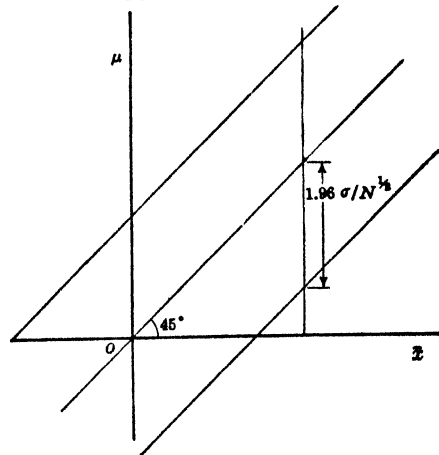


FIG. 45

If σ is not known it must be estimated from the sample. For large values of N , σ may be taken as the sample standard deviation s . For small N a

better estimate is $s \left(\frac{N}{N-1} \right)^{1/2}$, but unless σ is known, the equation (12.14) should not be used for small values of N , say less than 30 or 40. A better method for small samples is given in §12.10.

Example 2. The mean score for a randomly selected group of 50 students on an aptitude test was 493, with a standard deviation of 98. What are the 95% confidence limits for the mean score in the large population of students who took the test?

Here $\bar{x} = 493$, $s = 98$. Assuming that $\sigma = 98 \sqrt{50/49} = 99$, the confidence limits are $\mu = 493 \pm 1.96 \times 99/\sqrt{50} = 493 \pm 27$, so that when we fix the population mean, on the basis of this one sample, as lying between 466 and 520, we make a statement which, as one of many such statements, stands a 5% chance of being wrong.

12.5 Distribution of the Sample Sum. Instead of the mean \bar{x} we may be more interested in the sum $\sum x$ of all the x -values in the sample. Since $\sum x = N\bar{x}$, the distribution of $\sum x$ is of the same shape as that of \bar{x} , but with a mean and standard deviation N times as great. The mean of $\sum x$ is therefore $N\mu$ and the variance is $N\sigma^2$. Thus, in Example 1 of §12.3, the total number of spots in 50 throws of the die will have a mean $50 \times 3.5 = 175$ and a variance $50 \times 35/12 = 146$.

As an illustration of the distribution of the sum we shall give an alternative and simpler derivation of equations (10.15) and (10.16) for the mean and variance of the binomial distribution.

Consider a variable v which can take only two values 0 and 1, corresponding to "failure" and "success," and suppose the probability of 1 is θ . Then the probability of 0 is $1 - \theta$. For the distribution of v we have, therefore,

$$(12.15) \quad \begin{cases} \mu = (1 - \theta) \cdot 0 + (\theta) \cdot 1 = \theta \\ \sigma^2 + \mu^2 = (1 - \theta) \cdot 0^2 + (\theta) \cdot 1^2 = \theta \end{cases}$$

so that $\sigma^2 = \theta - \theta^2 = \theta(1 - \theta)$.

Now the sum of the values of v in s trials will be the number of successes in s trials, since each success contributes 1 to the value of v and each failure contributes 0.

The distribution of the number of successes x , in the binomial distribution, is therefore the same as the distribution of $\sum v$. By the result stated previously, the mean of $\sum v$ is s times the mean of v (that is, $s\theta$) and the variance is s times the variance of v (that is, $s\theta(1 - \theta)$), in agreement with the results deduced earlier. The same method can obviously be used to find higher moments of the binomial distribution, if these are required.

12.6 Correction for Finite Population. If a sample of size N is drawn from a population of size M , there is a finite number $C(M, N)$ of different samples which can be obtained. For example, if there are 20 numbered chips in a bowl and a sample of 5 is drawn, the number of different possible samples is 15,504. Each sample will have its own mean \bar{x} , given by adding

N values of x and dividing the sum by N . Let the mean of all the sample means be denoted by $\mu_{\bar{x}}$. Then the sum of all the sample means is $C(M, N)\mu_{\bar{x}}$, and the sum of all samples is $NC(M, N)\mu_{\bar{x}}$. But in adding all the possible samples every individual in the population must be included as often as every other, and since there are M individuals in the population and the total number of individuals in all samples is $NC(M, N)$, the number of times that each must be included is $NC(M, N)/M$. Therefore

$$NC(M, N)\mu_{\bar{x}} = \frac{NC(M, N)}{M} \cdot S$$

where S is the sum of all the individuals in the population and therefore is equal to $M\mu$. It follows that

$$(12.16) \quad \mu_{\bar{x}} = \mu$$

A similar but rather more complicated argument will show that

$$(12.17) \quad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{N} \cdot \frac{M - N}{M - 1}$$

so that the correction to the variance necessitated by the finite size of the population is the same as that given for the binomial distribution in §11.10.

For large values of M compared with N , the factor $(M - N)/(M - 1)$ is practically equal to 1, and then (12.17) reduces to (12.2). Thus, for the illustration in §12.2, dealing with samples of 10 from 1000 numbered discs, $(M - N)/(M - 1) = 0.991$, and the correction is therefore unimportant.

Example 3. In a freshman class of 180 students, the mean score on a test was 58, and the standard deviation was 12. If a group of 45 of these students is selected at random, what is the probability that the mean score for the group will be 60 or more?

Here we have $M = 180$, $N = 45$, $\mu = 58$, $\sigma = 12$, so that

$$\begin{aligned} \mu_{\bar{x}} &= 58 \\ \sigma_{\bar{x}} &= \frac{12}{(45)^{1/2}} \left(\frac{180 - 45}{180 - 1} \right)^{1/2} = 1.554 \end{aligned}$$

The z value corresponding to $\bar{x} = 60$ is therefore $z = \frac{60 - 58}{1.554} = 1.287$, and the probability of a value at least as great as this is 0.099.

12.7 Significance of Difference of Means in Large Samples. Suppose we have two samples of sizes N_1 and N_2 , with means \bar{x}_1 and \bar{x}_2 . If both samples are supposed drawn at random from the same population with mean μ and variance σ^2 , the difference $\bar{x}_1 - \bar{x}_2$ will be distributed with mean 0 and variance $\sigma^2/N_1 + \sigma^2/N_2$. This follows from Theorem 1 in §11.6. If both N_1 and N_2 are large enough for the means \bar{x}_1 and \bar{x}_2 to be practically normal (and for any N_1 and N_2 if the parent population is itself normal), the differ-

ence $\bar{x}_1 - \bar{x}_2$ will also be normal. That is to say, if

$$(12.18) \quad z = (\bar{x}_1 - \bar{x}_2)/\sigma \left(\frac{1}{N_1} + \frac{1}{N_2} \right)^{1/2}$$

z is a standard normal variate and the significance of a deviation from 0 can be assessed in the usual way. If the parent population is finite, of size M , σ^2/N_1 is replaced by $\frac{\sigma^2}{N_1} \cdot \frac{M - N_1}{M - 1}$ and σ^2/N_2 by $\frac{\sigma^2}{N_2} \cdot \frac{M - N_2}{M - 1}$.

Usually σ^2 is unknown and has to be estimated. An unbiased estimate of σ^2 is

$$(12.19) \quad \hat{\sigma}^2 = [N_1 s_1^2 + N_2 s_2^2]/(N_1 + N_2 - 2)$$

where s_1 and s_2 are the standard deviations of the samples.

For small samples the distribution of z when $\hat{\sigma}$ is substituted for σ is no longer normal, even for a normal parent population. The correct procedure is described in §12.11.

12.8 Student's t -distribution. We have seen that, for samples from a normal parent population, the quantity $z = \frac{\bar{x} - \mu}{\sigma/N^{1/2}}$ is a normal standard variate.

In most practical situations σ is unknown. If instead of σ we substitute the estimate $\hat{\sigma} = [N/(N - 1)]^{1/2} s$, where s is the standard deviation of the sample, then, as was first shown by W. S. Gosset, writing under the pen name of "Student" in a paper that has now become classic,* the variable

$t = \frac{\bar{x} - \mu}{s/(N - 1)^{1/2}}$ has a distribution which can be represented mathematically,

and for which tables like those of the normal law can be calculated. The numerator of t is normally distributed with mean zero, and the denominator is an *estimate*, from the sample, of the standard deviation of the numerator.

The important point about the distribution of t is that it depends only on the sample size N and not on the variance σ^2 of the population. In fact, the probability of a value of t between t and $t + dt$ is given by

$$(12.20) \quad f(t) dt = K \left(1 + \frac{t^2}{N - 1} \right)^{-N/2} dt$$

where K is a certain function of N . The curve of $f(t)$, plotted against t , is a symmetrical, hump-backed curve, not unlike the normal curve in shape but with higher tails. The curve for $N = 5$ is compared with the normal curve of equal variance in Fig. 46. As N increases the curve approaches the normal curve more and more closely. Its variance is $(N - 1)/(N - 3)$.

* See Reference 1. "Student" actually used the variable $(\bar{x} - \mu)/s$, which is not essentially different from t .

The probability of a value greater than a given value of t is

$$(12.21) \quad \int_t^{\infty} f(t) dt = 1 - F(t) = P$$

where $F(t)$ is the distribution function for t . Table II in the Appendix gives values of t corresponding to selected values of P for $n (= N - 1)$ between 1 and 30. The relation between t and P is illustrated in Fig. 46. The probability of a value at least as great *numerically* as the given t is double the probability stated. This double probability must be used in making a two-tailed test.

For values of n larger than 30, the quantity

$$z = t \left(\frac{n-2}{n} \right)^{1/2} = (\bar{x} - \mu)(n-2)^{1/2}/s$$

may be taken as approximately a standard normal variate

12.9 Degrees of Freedom. The quantity n used in Table II (Appendix) is called the number of *degrees of freedom*. The meaning of degrees of freedom is not easy to explain at an elementary level, but the general idea is that the N deviations $x_i - \bar{x}$ which are used in calculating the sample variance s^2 are not all independent, since there is a linear relation connecting them, namely,

$$\sum_{i=1}^N (x_i - \bar{x}) = 0$$

Because of this, only $N - 1$ of the quantities are actually independent, and we say that there are $N - 1$ degrees of freedom in the calculation of s . An unbiased estimate of the population variance is

$$Ns^2/(N-1) = \sum_{i=1}^N (x_i - \bar{x})^2/(N-1)$$

and is given by dividing the sum of squares of deviations from the mean by the number of degrees of freedom. For a good elementary discussion of the notion of degrees of freedom, the student may consult an article by Prof. Helen Walker (Reference 3). See also Part Two, page 162.

12.10 Confidence Limits for the Mean, for Small Samples. Given a sample, of size N , with mean \bar{x} and standard deviation s , we can use the

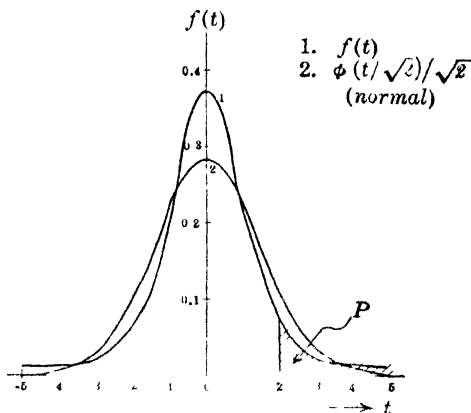


FIG. 46 STUDENT'S t ($n = 4$)

table of Student's t to find confidence limits for the mean μ of the population, on the assumption that x is normally distributed in this population. Writing n for $N - 1$, we have

$$(12.22) \quad t = (\bar{x} - \mu)n^{1/2}/s$$

For a given n we can find in the table the value of t corresponding to $P = 0.025$. The probability of a value of t numerically greater than this is 0.05, so that this value, substituted in (12.22) will give the 95% confidence limits for μ . Thus, if $n = 9$, $t = 2.262$, and we have

$$\frac{\bar{x} - \mu}{s} = \pm \frac{2.262}{\sqrt{9}}$$

or

$$(12.23) \quad \mu = \bar{x} \pm 0.754s$$

Example 4 For a sample of 10 rats, the mean blood viscosity reading was 3.93 and the standard deviation was 0.552. State 95% confidence limits for the blood viscosity reading in the population of rats sampled.

$$\text{From (12.23),} \quad \mu = 3.93 \pm 0.416$$

so that the confidence limits are 3.51 and 4.35. The constant in (12.23) is different for each different sample size. For large samples, it approximates $1.96/n^{1/2}$.

12.11 Confidence Limits for the Difference of Means, for Small Samples.

If we have two samples of sizes N_1 and N_2 , means \bar{x}_1 and \bar{x}_2 , and standard deviations s_1 and s_2 , we can form confidence limits for the difference of the means $\mu_1 - \mu_2$ in the two populations from which the samples are supposed to be taken, on the hypothesis that these two populations have the same variance. If the confidence limits include the value zero, we can say that the means of the two populations are not different, or, in other words, the hypothesis that the two samples come from populations with the same mean is not rejected at the chosen level.

If we denote the degrees of freedom for the two samples by n_1 and n_2 , we can prove that the quantity

$$(12.24) \quad t = [\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)]/\hat{\sigma}_{12}$$

where $\hat{\sigma}_{12}^2$ is an unbiased estimate of the variance of the numerator, is distributed as Student's t , with $n_1 + n_2$ degrees of freedom.

On the hypothesis of a common σ^2 , the variance of \bar{x}_1 is σ^2/N_1 , and that of \bar{x}_2 is σ^2/N_2 , so that the variance of $\bar{x}_1 - \bar{x}_2$ (the samples being independent) is

$$\sigma^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right) = \sigma^2(N_1 + N_2)/N_1N_2$$

The estimate of σ^2 is taken as a weighted mean of estimates from the two samples separately, weighted according to their respective degrees of freedom.

The estimate from the first sample is $N_1 s_1^2/n_1$ and that from the second is $N_2 s_2^2/n_2$, so that

$$(12.25) \quad \hat{\sigma}^2 = (N_1 s_1^2 + N_2 s_2^2)/(n_1 + n_2)$$

Then $\hat{\sigma}_{12}^2$ is given by

$$\begin{aligned} \hat{\sigma}_{12}^2 &= \hat{\sigma}^2(N_1 + N_2)/N_1 N_2 \\ &= \frac{N_1 s_1^2 + N_2 s_2^2}{n_1 + n_2} \cdot \frac{N_1 + N_2}{N_1 N_2} \end{aligned}$$

With this value of $\hat{\sigma}_{12}^2$, t in (12.24) has the Student t -distribution, and confidence limits can be found by putting t equal to the appropriate value corresponding to $n_1 + n_2$ degrees of freedom.

Example 5. Samples of two types of electric light bulbs were tested for length of life, and the following data were recorded:

Type 1	Type 2
$N_1 = 5$	$N_2 = 7$
$\bar{x}_1 = 1224$ hr	$\bar{x}_2 = 1036$ hr
$s_1 = 36$ hr	$s_2 = 40$ hr

Is the difference in the means sufficient to warrant the conclusion that Type 1 is superior to Type 2 in respect of length of life?

We assume that the standard deviation of life for the two types is the same. This may seem somewhat arbitrary, seeing that we do not suppose the means to be the same, but it may well happen that an improvement in construction increases the length of life of a lamp without much affecting its variability. Unless the assumption is approximately true, the t -test is not valid.

Here

$$\hat{\sigma}^2 = (5 \times 1296 + 7 \times 1600)/10 = 1768$$

$$\hat{\sigma}_{12}^2 = \frac{1768 \times 12}{35} = 606.2$$

and

$$\begin{aligned} t &= [1224 - 1036 - (\mu_1 - \mu_2)]/(\hat{\sigma}_{12}^2)^{1/2} \\ &= [188 - (\mu_1 - \mu_2)]/24.6 \end{aligned}$$

The number of degrees of freedom for t is $4 + 6 = 10$. The 95% value of t is 2.228, so that the 95% confidence limits are given by

$$\mu_1 - \mu_2 = 188 \pm 24.6 \times 2.228 = 188 \pm 54.8$$

and are therefore 133 and 243. Since these limits do not include zero, we can say that at the 5% level there is a significant difference between the means. If we feel sure that Type 1 cannot be worse than Type 2, but may be better, we shall use a one-tailed test, and the 95% value of t will be 1.812. The confidence limits for $\mu_1 - \mu_2$ are then

$$188 \pm 24.6 \times 1.812, \text{ or } 143 \text{ to } 233$$

If we want 99% confidence limits, then, for the two-tailed test, $t = \pm 3.169$, and the confidence limits are

$$\mu_1 - \mu_2 = 188 \pm 78$$

or 110 to 266. These limits still do not include zero, so that even at the 1% level the observed difference is significant.

12.12 Significance of Differences in Paired Samples. In some problems we are concerned with the effect of two different procedures or treatments carried out on the *same sample* of individuals. Instead of trying out one treatment on one random sample and the other treatment on an independent random sample, we try out both treatments on each member of the same sample. In this way we render the test more precise, because we eliminate a number of sources of variation which might possibly affect the quantity we are measuring. We ask now whether the mean of the individual differences between the x score on one treatment and that on the other treatment is significantly different from zero.

The null hypothesis is that there is no difference, in the population as a whole, between the x scores on the two treatments. We form a set of differences $d_i = x_{2i} - x_{1i}$, between the scores of the i th individual in the sample on the two treatments and assume that these differences are normally distributed in the population with mean zero. We then can use the t -test to find whether the *observed* mean of the d_i is significantly different from zero.

Example 6. Table 40 (*Smith and Medlicott*) shows the hemoglobin (gm/100 ml of blood) in anemic rats before and after 4 weeks of added iron in the diet (0.5 mg/day). The mean value of d is 0.825 and the standard deviation is 1.70. The number of degrees of freedom is 11 (one less than the number in the sample), so that, by (12.22),

$$t = (0.825 - 0)11^{1/2}/1.70 = 1.61$$

The probability of a value of t at least as great numerically as this, is between 0.1 and 0.2. The one-sided probability is between 0.05 and 0.1, and it would be reasonable to use a one-sided test on the principle that any difference in hemoglobin due to increased iron in the diet could only be an increase. However, even on this interpretation, the observed effect is non-significant.

TABLE 40. EFFECT OF IRON IN DIET ON HEMOGLOBIN

Rat No	x_1 (before)	x_2 (after)	$d = x_2 - x_1$
1	3.4	4.9	1.5
2	3.0	2.3	-0.7
3	3.0	3.1	0.1
4	3.4	2.1	-1.3
5	3.7	2.6	-1.1
6	4.0	3.8	-0.2
7	2.9	5.8	2.9
8	2.9	7.9	5.0
9	3.1	3.6	0.5
10	2.8	4.1	1.3
11	2.8	3.8	1.0
12	2.4	3.3	0.9

12.13 Distribution of the Sample Variance. It is proved in Part Two that the mean of the distribution of s^2 in samples of size N (from a parent population which is not necessarily normal) is equal to $(N - 1)\sigma^2/N$, where

σ^2 is the variance of the population. This fact has already been used in forming an estimate of σ^2 from an observed s^2 . It is also proved there that the variance of the distribution of s^2 is $\frac{N-1}{N^3} [(N-1)\mu_4 - (N-3)\mu_2^2]$, where μ_2 and μ_4 are the second and fourth moments for the population. If the distribution in the parent population is normal, $\mu_4 = 3\mu_2^2$, and the variance of s^2 then is given by

$$(12.26) \quad \text{Var}(s^2) = 2\sigma^4(N-1)/N^2$$

Usually σ^2 would be estimated by $Ns^2/(N-1)$ and the estimate of $\text{Var}(s^2)$ would then be

$$(12.27) \quad \text{Var}(s^2) = 2s^4/n$$

where $n = N - 1$. The square root of the estimated variance of a quantity is usually called its *standard error*.*

The distribution of s^2 is skew. It can be shown that when the parent population is normal the quantity Ns^2/σ^2 is distributed like a variate called χ^2 (Greek chi square) which is of great importance in a number of statistical problems. (The symbol χ^2 rather than χ is used because it is a quantity which can never be negative.) The graphs of χ^2 for two values of n are shown in Fig 47. The area of the righthand tail of the curve beyond a given

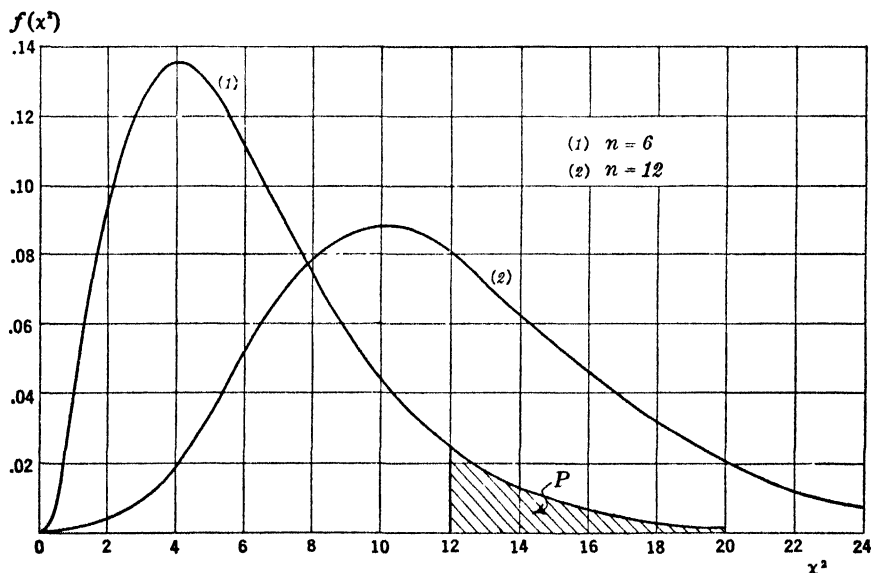


FIG. 47. DISTRIBUTION OF χ^2

* The term "probable error" occurs in the older literature for the standard error multiplied by 0.6745. If the distribution is normal, or nearly so, the probability is 0.5 that the variate lies within limits equal to the mean \pm the probable error (Cf. §8.5).

value of χ^2 is the probability P of a value of χ^2 greater than the given value. Table III in the Appendix gives for selected values of P the corresponding values of χ^2 , and from these the significance of an observed value can be judged. The table of χ^2 enables us also to set confidence limits for σ^2 from the observed value of s^2 in a sample. Thus for 90% confidence limits, we find from the table the two values of χ^2 which correspond to $P = 0.95$ and $P = 0.05$. For $N = 10$ (and therefore $n = 9$) the two values of χ^2 are 3.325 and 16.919. The confidence limits for σ^2 are given by putting $10s^2/\sigma^2$ equal to these two values, and so are $10s^2/3.325$ and $10s^2/16.919$, that is, $3.01s^2$ and $0.59s^2$.

12.14 Significance of a Ratio of Two Variances. The F -distribution.

Suppose we have two independent random samples of sizes N_1 and N_2 and variances s_1^2 and s_2^2 . We may wish to test whether the null hypothesis, that is, that the two samples come from populations with the same variance σ^2 , is justified. We make this assumption, for instance, in testing the significance of the difference of the means, and it would be satisfactory, before going ahead with the t -test, to be assured that the assumption is a reasonable one.

It turns out that instead of using the difference of the two variances it is more convenient to use their *ratio*. We find that the quantity

$$(12.28) \quad F = \frac{N_1 s_1^2}{n_1} \bigg/ \frac{N_2 s_2^2}{n_2}$$

which is a ratio of the estimate of σ_1^2 from the first sample to the estimate of σ_2^2 from the second, has a distribution which, on the hypothesis that the two samples are from the same population with variance σ^2 , is independent of σ^2 and depends only on the two numbers n_1 and n_2 . As would be expected, F fluctuates around the value 1; the mean of the distribution is $n_2/(n_2 - 2)$, for $n_2 > 2$. Tables of the area under the curve of $f(F)$ beyond a given value of F enable us to find the probability P that F will exceed the observed value, and hence to judge the significance of an observed ratio of s_1^2 to s_2^2 . Such tables are given in Table IV of the Appendix. Since a complete table, giving P for all reasonable values of F , n_1 and n_2 , would be very bulky, the table gives only the values of F corresponding to *two* selected values of P , namely, 0.05 and 0.01. These are called the 5% points and the 1% points, respectively.

The null hypothesis is that the ratio of σ_1^2 to σ_2^2 is equal to 1. If we are going to reject this hypothesis both when F is too great and when F is too small, the 5% point really corresponds to a 10% level of significance, since we are using a two-tailed test, and each tail has an area of 5%. If, on the other hand, the alternative hypothesis is that $\sigma_1^2/\sigma_2^2 > 1$, we use a one-tailed test and then the 5% point corresponds to a 5% level of significance. This is the usual situation in problems of Analysis of Variance, which provide the main occasions for using Table IV.

Example 7. In two samples, of sizes 6 and 11, the observed standard deviations of a variable x are 3.6 and 2.0. Is this difference significant?

$$\text{Here} \quad F = \frac{6(3.6)^2}{5} \cdot \frac{10}{11(2.0)^2} = 3.53$$

with $n_1 = 5$, $n_2 = 10$.

From the table we see that the 5% point is 3.33 and the 1% point is 5.64. The difference is therefore significant at a level a little below 10%, if we have no reason to believe that either population has the greater variance. If we feel sure that, provided there is any difference, the population from which sample 1 is taken will have the larger variance, the observed difference is significant at a level below 5%.

Table IV gives the F values at the stated levels for the upper tail of the distribution. Values for the lower tail can be obtained, if desired, by taking the reciprocal of F , at the same time interchanging the degrees of freedom n_1 and n_2 . Thus the upper 5% point for $n_1 = 5$, $n_2 = 10$, is 3.33, and the lower 5% point is $1/4.74 = 0.211$. Here 4.74 is the tabulated point for $n_1 = 10$ and $n_2 = 5$. However, in practice we seldom need the lower tail, as we can always arrange the ratio F so that the estimate in the numerator is greater than that in the denominator.

12.15 Test for Homogeneity of Variance. If we have several samples we can test the null hypothesis that they all come from populations with the same variance. Consider, for example, the data in Table 41, on six samples, each of five items, where the variable is the breaking strength in lb wt of a specimen of cotton cloth. For each sample an estimate of the population variance σ^2 can be calculated, and these estimates, denoted by $\hat{\sigma}_k^2$, are given in the last row but one of Table 41. They evidently differ quite considerably.

TABLE 41. BREAKING STRENGTH OF COTTON CLOTH (LB WT)

Sample No.	1	2	3	4	5	6
	38.6	47.4	47.7	44.6	48.1	45.2
	46.5	42.3	44.7	49.8	50.6	42.1
Items	43.1	46.3	46.4	42.1	41.3	51.5
	47.2	46.0	46.0	47.3	41.3	50.6
	53.4	48.9	46.4	52.2	48.1	46.4
\bar{x}_k	45.76	46.18	46.24	47.20	45.88	47.16
$\hat{\sigma}_k^2$	29.83	6.00	1.15	16.14	18.52	15.17
$\log_{10} \hat{\sigma}_k^2$	1.475	0.778	0.062	1.208	1.268	1.181

Now it can be shown that if the null hypothesis is true, and if n_k is the number of degrees of freedom for the k th sample, then the quantity M defined by

$$(12.29) \quad M = n \log \left(\sum_k n_k \hat{\sigma}_k^2 / n \right) - \sum_k n_k \log \hat{\sigma}_k^2$$

where $n = \sum n_k$, is independent of σ^2 and can be used to find the probability of obtaining the observed set of variances if the estimate of σ^2 from the whole

set of samples is fixed, that is, for a fixed value of $S/n = \sum n_k \hat{\sigma}_k^2 / n$. The greater the value of M , the smaller this probability, and hence the smaller the chance of obtaining the heterogeneous collection of variances actually observed if the samples form a set of random samples from the same population. The logarithms in (12.29) are natural logarithms. If common logs are used, M must be multiplied by 2.303.

Bartlett proved that, if b is the number of samples, and if the n_k are reasonably large, M is approximately distributed like χ^2 with $b - 1$ degrees of freedom, so that the table of χ^2 can be used to test the significance of the differences between the $\hat{\sigma}_k^2$. For very small samples, down to a size of 4 or 5, it is better to use M/c instead of M where c is a correction factor given by

$$(12.30) \quad c = 1 + \{ \sum (1/n_k) - 1/n \} / 3(b - 1)$$

If all the n_k are equal, $n_k = n/b$, so that

$$(12.31) \quad \begin{cases} M = 2.303 n [\log_{10} (\sum \hat{\sigma}_k^2 / b) - (\sum \log_{10} \hat{\sigma}_k^2) / b], \\ c = 1 + (b + 1) / 3n \end{cases}$$

For the data of Table 41, $n_k = 4$, $b = 6$, $n = 24$, $c = 1 + 7/72 = 1.097$.

Also,

$$M = 55.26 [\log_{10}(86.81/6) - (5.971)/6] = 9.13$$

so that $M/c = 8.33$. For 5 degrees of freedom, the 10% point for the distribution of χ^2 is 9.236, so that there is a probability of more than 0.1 of obtaining a set of variances as unlike as the $\hat{\sigma}_k^2$ on the null hypothesis. The null hypothesis can therefore be accepted quite reasonably.

12.16 Analysis of Variance. We can look at the data of Table 41 in a different way, asking the question whether the sample means differ among themselves more than would be expected from the way in which the individual items in a single sample differ. If we regard the whole set of $\sum (n_k + 1) = n + b$ items as a single sample, we can form the mean \bar{x} of all these items and also the sum S_t of all the squared deviations from the mean, $(x_{kj} - \bar{x})^2$, where x_{kj} is the j th item in the k th sample. S_t is called the "total sum of squares," and, on the null hypothesis that all the samples came from populations with the same mean and the same variance σ^2 , $S_t/(n + b - 1)$ is an estimate of σ^2 with $n + b - 1$ degrees of freedom. A second estimate of σ^2 is obtained by averaging the separate estimates $\hat{\sigma}_k^2$ from the b separate samples, the average being weighted according to the degrees of freedom n_k . This average is the quantity $\sum n_k \hat{\sigma}_k^2 / n$ denoted by S/n in the preceding section. Since $n_k \hat{\sigma}_k^2 = \sum_j (x_{kj} - \bar{x}_k)^2$, S is defined as $\sum_k \sum_j (x_{kj} - \bar{x}_k)^2$. It is called the "sum of squares within samples," being a measure of the variation within each sample around its own mean. Finally, we can form a third estimate of σ^2 from the variation between the separate means \bar{x}_k . The variance of the

mean of a sample of size $n_k + 1$ is equal to the variance of the sample itself divided by $n_k + 1$, so that if $S_b = \sum (n_k + 1)(\bar{x}_k - \bar{x})^2$, the quantity $S_b/(b - 1)$ is an estimate of σ^2 . S_b is called the "sum of squares between samples," being a measure of the variation between the means of the different samples. The estimates $S_b/(b - 1)$ and S/n are independent, and if we make the further assumption that the variable x is normally distributed in the populations from which the samples are taken, we can say that the ratio

$$(12.32) \quad F = \frac{S_b}{S} \cdot \frac{n}{b - 1}$$

has the F -distribution of §12.14. The tables of F can then be used to determine whether the null hypothesis is justified. If the null hypothesis is rejected, because the F value is too large, the conclusion is that the differences between the means are significant. Often the samples have been subjected to different treatments, or have been taken at different times, and if so we are enabled to say whether or not the treatments or the lapse of time has produced a definite effect on the variate x . The three estimates of σ^2 can be set out in a table (see Table 42) giving the respective sums of squares and the corresponding degrees of freedom. Since the total variance is analyzed into a part due to variation *between* the samples and a part due to variation *within* the samples, this process is known as *Analysis of Variance*. It is a very powerful tool in statistical investigation, and can be used in much more complicated situations than the one we have envisaged here. For further details the student may consult the chapter on the subject in Part Two, or the textbook by Snedecor (Reference 14 of §0.4).

It can be shown mathematically that S_t is the sum of S and S_b . Also the number of degrees of freedom for S_t is clearly the sum of the degrees of freedom for the other two. For the purpose of calculation it is preferable to put the expressions for the sums of squares in the form:

$$(12.33) \quad \begin{cases} S_t = \sum_{jk} (x_{kj})^2 - (\sum_{jk} x_{kj})^2 / N \\ S_b = \sum_k (\sum_j x_{kj})^2 / N_k - (\sum_{jk} x_{kj})^2 / N \\ S = S_t - S_b \end{cases}$$

where $N = n + b$ = total number in all the samples, and N_k is the number in the k th sample.

As applied to the data of Table 41, we find the values for these sums of squares given in Table 43, and it is clear that there is no significant difference between the means of the separate samples. In fact, these means are much closer together than we should expect, even if they were random samples from the same population. If it had turned out that the "between" estimate of

TABLE 42. ANALYSIS OF VARIANCE

<i>Variation</i>	<i>Sum of Squares of Deviations</i>	<i>Degrees of Freedom</i>	<i>Estimate of σ^2</i>
Within Samples	$S = \sum_{jk} (x_{kj} - \bar{x}_k)^2$	n	S/n
Between Samples	$S_b = \sum_k (n_k + 1)(\bar{x}_k - \bar{x})^2$	$b - 1$	$S_b/(b - 1)$
Total	$S_t = \sum_{jk} (x_{kj} - \bar{x})^2$	$n + b - 1$	$S_t/(n + b - 1)$

TABLE 43. ANALYSIS OF VARIANCE OF DATA IN TABLE 41

<i>Variation</i>	<i>Sum of Squares</i>	<i>Degrees of Freedom</i>	<i>Estimate of σ^2</i>
Within Samples	347.25	24	14.5
Between Samples	9.86	5	1.97
Total	357.11	29	12.3

σ^2 was about 2.6 times as great as the "within" estimate, we could have said that there is a barely significant difference between sample means at the 5% level, the 5% value for $n_1 = 5$ and $n_2 = 24$ being 2.62.

12.17 Control Charts. Data like those in Table 41 on the breaking strength of cotton cloth are often obtained as a routine in factories where products are manufactured to a specification, or where they have to conform to buyers' standards. The manufacturer of cloth expects a certain amount of variation in the strength of the product, but he wishes the average strength to be maintained and the variation around the average to be kept within reasonable bounds. To ensure this, samples are taken regularly and the results of measurements on these samples are plotted on *control charts*, so that any unusual features may be readily noted and action taken if necessary to bring the process back under control. The usual control charts are for the sample mean and the sample range, the latter being used as a quick and easy estimate of dispersion. Sometimes sample inspection involves, instead of a measurement, merely the decision as to whether a manufactured article is defective or not. The proportion of defective articles in the sample can be similarly set out in a control chart.

A typical control chart for means is shown in Fig. 48. The central line and the upper and lower lines are placed on the chart after a fairly long sequence of readings has shown that the process is behaving in a reasonably steady way. The central line is drawn at a point on the vertical scale which represents an estimate μ of the population mean based on 30 or 40 samples. The other lines are upper and lower limits for \bar{x} based on the estimated sampling standard deviation of the mean, $\hat{\sigma}_{\bar{x}}$, and so placed that the probability

is very small that a value of \bar{x} will, by pure chance, fall outside these limits. How small is a matter of choice. It is customary to place the limits at $\mu \pm 3\hat{\sigma}_{\bar{x}}$, and the probability is then only about 0.003, but we could, of course, choose 0.01 or 0.05. The reason for making the limits wide is that the manufacturer does not wish to waste time looking for non-existent trouble, and the chance that he will do so with the $3\hat{\sigma}_{\bar{x}}$ limits is only about 3 in 1000. Of

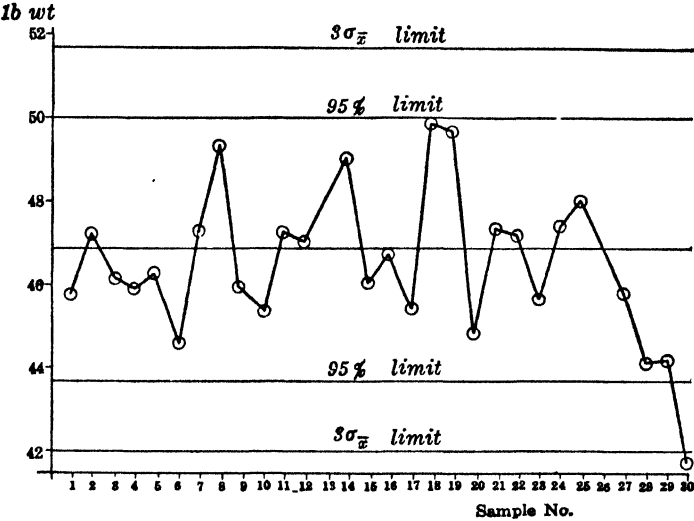


FIG. 48. CONTROL CHART FOR MEANS

course, there is a correspondingly greater chance of failing to recognize trouble that really exists. The danger signals in a control chart are (1) plotted points falling outside the control limits, (2) several consecutive points lying near the limits, although still inside, (3) runs of 7 or more points all above or all below the central value, or (4) a well-marked trend upward or downward in the plotted points. If conditions (3) or (4) are observed it may be that new control limits need to be calculated, as the process is now settling

TABLE 44. CONTROL CHART LIMITS FOR RANGE

Sample Size	99% values of R/\bar{R}		3σ values of R/\bar{R}	
	Lower	Upper	Lower	Upper
4	0.165	2.28	0	2.282
5	0.237	2.10	0	2.114
6	0.296	1.99	0	2.004
7	0.340	1.90	0.076	1.924
10	0.432	1.76	0.223	1.777
15	0.524	1.64	0.348	1.652

down to a state different from that in which the original control limits were calculated.

In a control chart for the range, the $3\sigma_R$ limits and also the 99% limits are given for a few sample sizes in Table 44, in terms of the mean \bar{R} obtained from the preliminary set of samples. The values are based on the assumption that \bar{R} is the true mean range for all possible samples of the given size selected from the population, which is supposed to be normal.

For further information on quality control, see Reference 7 of §0.4, or Reference 2 of Chapter IV.

Exercises

1. A normal population has mean 20 and standard deviation 2. A sample of 6 items from this population has a mean 18.2. Can the sample be reasonably regarded as a random sample from the population? *Ans.* Not at the 5% level.

2. A very large population has a mean 26.54 ft and a variance 28.3 ft². What percentage of random samples of size 135, taken from this population, will have means differing from the population mean by more than 1 ft? What percentage of samples of size 200 will have means between 26.1 and 26.7 ft? *Ans.* 2.9%, 54.4%.

3. A population is known to be normal and to have a standard deviation of 0.104 second. A random sample of 12 items has a mean of 12.33 seconds. Calculate 95% confidence limits for the population mean.

4. For the population of Exercise 3, what is the smallest sample size that will ensure, with 95% confidence, that the sample mean will not differ from the population mean by more than 0.05 second? *Ans.* 17.

5. A population is known to have a mean of about 25 bushels/acre and a coefficient of variation of about $\frac{1}{4}$. You wish to find, from a single random sample, 95% confidence limits for the population mean. If these limits are not to differ from each other by more than 1 bushel/acre how large should the sample be? *Ans.* 600.

6. Four different boxes of Eddy's matches, from the same carton, contained 55, 58, 53, and 57 matches. Obtain 95% confidence limits for the mean number of matches in a box.

7. A group of 120 freshmen at University A take a certain standard test and obtain a mean score 70 with a standard deviation 14. A group of 80 freshmen at University B take the same test and obtain a mean score of 75 with standard deviation 12. Test the hypothesis that the two groups are random samples from the same population (that is, that the difference between the mean scores is not significant).

8. Twenty lines, each exactly 6 inches long, were drawn. A student estimated by eye the center of each line. The distance in inches of each point, so estimated, from the left-hand end of the line was measured, with the following results: 2.97, 3.11, 2.97, 3.18, 3.13, 3.23, 2.98, 3.02, 2.92, 3.13, 3.00, 3.06, 2.98, 3.00, 2.94, 2.98, 3.07, 3.07, 3.03, 3.20. Obtain 95% confidence limits for the population mean. What is the population in this example? Is there any reason to believe that the student was making a systematic error?

9. (*Wilks*) An aptitude test was given to two groups of soldiers: (a) a group of 1050 who had been in the army for some time; (b) a group of 531 new selectees. The means and standard deviations of scores for the two groups were as follows:

	\bar{x}	s_x
(a)	47.65	6.77
(b)	46.10	6.79

Find 95% confidence limits for the difference between the population means.

10. Two distinct groups of rats, (a) normal and (b) adrenalectomized, were tested for blood viscosity. The sample sizes and the means and standard deviations of viscosity readings were:

	N	\bar{x}	s_x
(a)	11	3.921	0.527
(b)	9	4.111	0.601

Is the difference in the means significant at the 5% level?

Ans. No. The 95% confidence limits for $\mu_1 - \mu_2$ are 0.19 ± 0.56 .

11. (*Snedecor*) (a) An agronomist interested in the effect of superphosphate on corn yield added the fertilizer to a mixture of manure and lime. The old and new fertilizers were tested on five pairs of adjacent plots, the plots in each pair being as alike as possible except that one member was treated with the old fertilizer and one with the new. The plots with the superphosphate yielded 20, 6, 4, 3, and 2 bushels per acre more than their parallels. Was the value of the superphosphate demonstrated?

(b) Suppose the increased yields had been 5, 6, 4, 3 and 2 bushels per acre. Would the value of the superphosphate have been demonstrated then? Explain the apparent paradox.

12. A physiological experiment was carried out to test the effect of an injection of secretin on the percentage of reticulocytes in the blood of rabbits. 17 rabbits were tested before and after injection, the mean of the increases was 0.0635, and the standard deviation of increases was 0.168. Was the existence of an effect demonstrated?

13. The densities of sulphuric acid in two containers were measured, four determinations being made on one and six on the other. The results were:

(1)	1.842,	1.846,	1.843,	1.843		
(2)	1.848,	1.843,	1.846,	1.847,	1.847,	1.845

Is the difference significant at the 5% level,

(a) if there is no reason beforehand to believe that either lot of acid is the denser?

(b) if we have good reason to suspect that if there is any difference the first will be lighter than the second?

Ans. (a) no; (b) yes.

14. The following table gives the strength (lb wt/in.²) of concrete made with sand containing different percentages of coal. Each sample of concrete was made into four cylinders which were tested for strength.

Sample number	Percentage coal	x (lb wt/in. ²)			
1	0.00	1690,	1580,	1745,	1685
2	0.05	1550,	1445,	1645,	1545
3	0.10	1625,	1450,	1510,	1590
4	0.50	1725,	1550,	1430,	1445
5	1.00	1530,	1545,	1565,	1520

Test the homogeneity of the variances of these samples. Also test the means of samples 2, 3, 4, and 5, each compared with the mean of sample 1.

15. Arrange the data of Exercise 14 as an analysis of variance, and use the *F* test to determine whether the differences between sample means are significant as compared with the differences within the samples.

16. Five samples, each of four seasoned mine-props, were tested for maximum load. The means and standard deviations of the maximum load (in units of 1000 lb wt) were

<i>Sample No.</i>	\bar{x}	s_x
1	42.0	8.75
2	52.0	10.44
3	65.5	4.72
4	51.8	8.26
5	73.5	16.68

For each sample estimate 95% confidence limits for the mean, and do the same for the combined sample of 20. Test the homogeneity of the variances.

17. Two small samples of herring were measured for length (mm), with the following results:

(1) 192, 179, 181, 193, 215, 181, 178

(2) 173, 194, 194, 187, 168, 186, 176, 191, 191, 178, 185, 160

Do these samples differ significantly in average length?

18. A manufacturer desires to turn out cotton thread, the breaking strength of which is to have a mean and standard deviation 6.50 oz and 1.50 oz, respectively. Assuming that this standard has been attained, what should now be the 99% and the 3σ control limits for the mean of routine samples of 10 pieces of thread?

19. Use the method of the auxiliary variable v in §12.5 to find the skewness and kurtosis of the binomial distribution.

References

1. "Student," "The Probable Error of the Mean," *Biometrika*, **6**, 1908, pp. 1-25. For an interesting biographical sketch of Mr. W. S. Gosset, see Reference 2.
2. *J. Amer. Stat. Assoc.*, **33**, 1938, pp. 226-228.
3. H. M. Walker, *Journal of Educational Psychology*, **31**, 1940, pp. 253-269.

CHAPTER XIII

NON-PARAMETRIC AND ORDER STATISTICS

13.1 Non-parametric Statistics. In the problems of estimation we have so far encountered, we have assumed the *form* of the distribution of our variate in the parent population (for example, binomial or normal) and have endeavored to find best values and confidence limits for one or more *parameters* of this distribution. Sometimes, however, we wish to infer something about the distribution as a whole and are not concerned with the numerical values of the parameters. Problems of this nature are called *non-parametric*.

13.2 Goodness of Fit. In Chapters VIII and X we saw that some empirical distributions arising out of sampling experiments, and some observed frequency distributions, can be fitted more or less closely by theoretical distributions of the binomial, Poisson, or normal types. No criterion of the goodness of fit was mentioned, however. We shall now discuss a widely used technique, known as the *Chi-square* (χ^2) *test*, for judging whether or not the fit is satisfactory.

Consider, for example, the binomial distribution fitted to the result of a sampling experiment in Table 35, §10.5. For each value of x from 0 to 10 we have in the second column an observed frequency (f_0) and in the third column a calculated frequency (f_c), based on an assumed probability of success (θ) equal to $\frac{1}{3}$. The values f_c , divided by the total frequency N , represent the *probabilities* (on the hypothesis that $\theta = \frac{1}{3}$) of exactly 0, 1, 2, \dots , 10 successes in a trial, "success" meaning here a red ball in a sample of 10 balls. On the basis of these probabilities we can calculate the chance of getting, in 350 samples, the observed set of frequencies f_0 , that is, the chance that there will be 2 cases with $x = 0$, 22 with $x = 1$, and so on. This distribution in *classes* is called a *multinomial distribution* (the binomial is a special case with only two classes). It was proved by Karl Pearson that the quantity

$$(13.1) \quad \chi^2 = \sum (f_0 - f_c)^2 / f_c$$

where the sum is over the k classes in the distribution, is, for large values of N , distributed like the quantity χ^2 described in §12.13, with $k - 1$ degrees of freedom. For any given set of frequencies f_0 and a corresponding set of independently calculated frequencies f_c , we can therefore find the value of χ^2 and use the table of χ^2 to test its significance.

Clearly, the greater the differences between the observed and calculated frequencies, the greater will be the value of χ^2 , so that, generally speaking, the larger χ^2 , the worse the fit. More precisely, the area under the χ^2 curve,

beyond the given value of χ_s^2 , represents the probability of obtaining by chance, on the null hypothesis that the distribution in the parent population is the assumed theoretical one, a value of χ_s^2 at least as great as the one actually found, or in other words the probability of a fit at least as bad as the observed fit. If this probability is not too small, the null hypothesis may be accepted and the fit regarded as satisfactory. If, for example, there are 10 classes in a distribution (so that $k - 1 = 9$), we see from Table III in the Appendix that the probability that $\chi^2 > 16.919$ is 0.05. A value of χ_s^2 of 16 or less would be regarded by most statisticians as not furnishing good evidence *against* the null hypothesis, and the null hypothesis could therefore reasonably be accepted. On the other hand, the probability is less than 0.01 that $\chi^2 > 22$ (for 9 degrees of freedom) and a value of χ_s^2 as large as this *would* provide good evidence against the null hypothesis. Since the test of goodness of fit is concerned with the whole course of the distribution, it is a non-parametric test.

13.3 Pooling of Class Frequencies. The proof of the approximate χ^2 distribution for χ_s^2 in (13.1) rests on the assumption that none of the frequencies f_c is very small. How small f_c may safely be is an open question, but we shall probably not be far wrong in putting 5 as the lower limit. If it happens (as it often does in practice) that a few of the end classes have very small frequencies, it is well to group these classes together until no class contains fewer than 5, before applying the χ^2 test. Thus in the data of Table 35, §10.5, the classes $x = 8, 9$, and 10 have respectively $f_c = 1.1, 0.1$, and 0.0, so that these should be pooled with the class $x = 7$. The number of classes in the table is then reduced to 8, with 7 degrees of freedom. The loss of one degree of freedom is here attributable to the fact that the total frequency is fixed. In distributing 350 objects among 8 classes we can (within limits) put as many as we like in 7 of the 8 classes, but the number in the 8th class is then determined by the number of objects we have left.

The calculation of χ_s^2 for the data of Table 35 with $\theta = 1/3$ is shown in Table 45. The sum of the last column is 15.2, and the probability of a value

TABLE 45. GOODNESS OF FIT OF BINOMIAL

X	f_0	f_c	$f_0 - f_c$	$(f_0 - f_c)^2$	$(f_0 - f_c)^2/f_c$
0	2	6.1	-4.1	16.8	2.8
1	22	30.3	-8.3	68.9	2.3
2	63	68.3	-5.3	28.1	0.4
3	76	91.0	-15.0	225.0	2.5
4	96	79.7	16.3	265.7	3.3
5	56	47.8	8.2	67.2	1.4
6	26	19.9	6.1	37.2	1.9
7-10	9	6.9	2.1	4.4	0.6
	<u>350</u>	<u>350.0</u>			<u>15.2</u>

of χ^2 at least as great as this is about 0.033. The 5% point is 14.1 and the 1% point 18.5. Our value therefore is great enough to make us reject the hypothesis that the distribution is really binomial with parameter $\theta = \frac{1}{2}$.

13.4 The Chi-Square Test of Hypotheses. A particular distribution among classes may be suggested by a preconceived theory or hypothesis, and this theory can then be tested by comparing the suggested distribution with the one actually observed. Thus, in a classic experiment, the Abbe Mendel observed the shape and color of peas from a number of plants in the first generation progeny of a cross, and found that they could be classified in four groups, as follows:

Round, yellow	315
Round, green.....	108
Angular (wrinkled), yellow.....	101
Angular, green.....	32

According to Mendel's theory of heredity the expected numbers should be in the ratio 9 : 3 : 3 : 1, and therefore, for a total of 556, are 312.75, 104.25, 104.25, and 34.75. The value of χ^2 from these data is 0.47, with 3 degrees of freedom. The probability of a value of χ^2 at least as great as this is about 0.92, so that the agreement of theory and experiment is closer than would be expected.

Very high values of the probability (say higher than 0.99) are sometimes encountered and are usually to be viewed with suspicion. The fit is too good to be true, and it is likely that the sample investigated is not truly a random sample of the population.

13.5 The Chi-Square Test of Goodness of Fit for Graduated Distributions.

In fitting a theoretical curve to a distribution* it is often necessary to calculate one or more parameters of the curve from the distribution itself. The true values of the parameters are therefore replaced by estimates, and it can no longer be taken for granted that the limiting distribution of χ^2 for large N is a χ^2 distribution. However, if the estimates are the best possible ones (said to be *most-efficient*) it is true that the limiting distribution is a χ^2 distribution with $k - 1 - p$ degrees of freedom, where p is the number of parameters estimated from the sample. It is unfortunately not true that the method of moments (the one we have adopted in earlier chapters) always gives most-efficient estimates, but if the theoretical distribution is normal, binomial, or Poisson this method is quite satisfactory.

The number p of parameters varies with the type of distribution, being 1 for the Poisson curve, 2 for the normal, and 3 or more for various skew curves. For the *binomial* there is usually only one parameter (θ) to estimate, the other (s) being fixed by the conditions of the problem. In Table 35, §10.5, values

* The curve is, of course, "fitted" to a *histogram*. The term "theoretical curve" is often used to mean a theoretical distribution, particularly for a continuous variate.

of f_c are calculated in column 4 for a value of θ (0.36) estimated from the sample. If we pool the frequencies for $x = 0$ and 1 and also for $x = 7, 8, 9$, and 10, we have 7 classes and the number of degrees of freedom is 5. The value of χ_s^2 turns out to be 3.81, and the probability of a fit worse than this is about 0.57. The fit is therefore excellent.

In fitting a *normal curve* to the distribution of weights of Glasgow school-girls (see Fig. 31, §§8.6 and 8.7), two parameters, μ and σ , were estimated from the sample. In Table 31 theoretical frequencies are given for each class interval corresponding to the observed frequencies for the same intervals. This being a continuous distribution, the theoretical frequencies are areas under the normal curve between ordinates erected at the class boundaries; the first area extends from $-\infty$ to the end of the first interval and the last area extends from the beginning of the last interval to $+\infty$. Thus, for the interval from 35.5 to 39.5 lb, $f_c = 60.3$. This area, divided by the total frequency, is denoted by ΔA ($= 0.0603$) and is, on the null hypothesis, the *probability* that a schoolgirl selected at random out of the whole population of Glasgow (in the age group of the sample and at the time the sample was taken) would have a weight between 35.5 and 39.5 lb.

If the two classes at the beginning and the two classes at the end in Table 31 are pooled, we have the comparison of observed and calculated frequencies given in Table 46. It is evident that the fit is excellent.

TABLE 46. GOODNESS OF FIT OF NORMAL CURVE

f_o	f_c	$f_o - f_c$	$(f_o - f_c)^2$	$(f_o - f_c)^2/f_c$
15	17.2	-2.2	4.8	0.28
56	60.3	-4.3	18.5	0.31
172	155.5	16.5	272.2	1.75
245	252.4	-7.4	54.8	0.22
263	258.7	4.3	18.5	0.07
156	167.2	-11.2	125.4	0.75
67	68.1	-1.1	1.2	0.02
26	20.6	5.4	29.2	1.42
1000	1000.0	0		4.82

$$\chi_s^2 = 4.82, \quad n = 8 - 1 - 2 = 5$$

$$P = 0.45$$

13.6 Tests of Randomness. We have repeatedly used the word "random" in speaking of samples and have defined a random sample as one in which every individual in the population has an equal chance of being included. This definition is not, however, of much use in deciding practically whether a sample is random or not. We have to look at the various items of the sample, *in the order in which they come*, and see whether or not they exhibit a satisfactory degree of haphazardness. In calculating the statistics of a

sample we have not hitherto troubled about the order of arrangement of the items in the sample, but this order is essential in a discussion of randomness.

Consider, for simplicity, the tossing of a coin, and suppose we denote "head" by 0 and "tail" by 1. Then a succession of tosses, in the order in which they are made, will be represented by a set of numbers like

(13.2) 0 1 1 0 0 0 1 0 1 0 0 1 1 1 1 0 1 1 0 1 0 1 0 1 0 0...

We naturally expect that the proportion of 1's in such a set will tend, as the number of tosses increases, toward a value in the neighborhood of $\frac{1}{2}$. But, apart from this, we also expect the sequence to be haphazard. That is, we should think it very strange to get a sequence like

(13.3) 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1...

or

(13.4) 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 1 1...

One method of testing a very long sequence for randomness is to pick out a subsequence (for example, every second term) and see whether the proportion of 1's in the subsequence is the same as in the original sequence. If it is so in *every* subsequence that we can choose in this way, then, according to R. Von Mises (see Reference 1, page 143), the sequence is random, but this is clearly not a practicable test, although we can, of course, try it on a *few* subsequences. Thus, if we choose the 1st, 3rd, 5th . . . items of (13.2) we get 0 1 0 1 1 0 1 1 1 0 0 0 0 . . . , and the proportions of 1's is still near $\frac{1}{2}$, but if we do the same in (13.3) we get only 0's. This test, therefore, rules out (13.3) as random. It does not rule out (13.4), but if we take a different subsequence (choosing 1st, 13th, 25th, etc., at intervals of 12), we get all 1's, and thus (13.4) is also non-random by the foregoing test.

A more practical test of randomness is based on the number of *runs* of zeros and ones in the sequence. A run of zeros is a set of successive zeros closed off at both ends by 1's (except at the beginning and end of the sequence) and similarly for a run of 1's. Thus in (13.2) we have in succession runs of one 0, two 1's, three 0's, one 1, one 0, one 1, two 0's and so on, a total of 17 runs in all. In (13.3) there are 26 runs and in (13.4) only 5 runs. Clearly, a non-random sequence may have a large number of runs or a small number, compared with a random sequence. The probability of an assigned number of runs, on the hypothesis of randomness, can be calculated, and also upper and lower limits can be assigned within which we can be confident (with a specified degree of confidence) that the number of runs will lie. If it does not, we reject the hypothesis of randomness at the corresponding level. It turns out that in a sequence of 13 0's and 13 1's (like (13.2)), the number of runs should, at the 95% level, lie between 9 and 19 inclusive, so that this test rejects (13.3) and (13.4) without rejecting (13.2).

13.7 Distribution of Number of Runs. If we have m objects of one kind (say 0's) and n objects of another kind (say 1's) arranged along a line, the number of possible arrangements is $C(m+n, m)$. We will suppose that $m \leq n$ (that is, we let m refer to whichever set of objects is the fewer in number). The number of arrangements with exactly u runs can be shown to be

$$(13.5) \quad f_u = 2C(m-1, k-1)C(n-1, k-1)$$

if u is even ($= 2k$) and

$$(13.6) \quad \begin{aligned} f_u &= C(m-1, k-1)C(n-1, k-2) \\ &\quad + C(m-1, k-2)C(n-1, k-1) \end{aligned}$$

if u is odd ($= 2k-1$). Here k can take all integral values from 1 to $m+1$. The probability that the number of runs is equal to or less than u' in a random arrangement is therefore given by

$$(13.7) \quad P\{u \leq u'\} = \sum_{u=2}^{u'} f_u / C(m+n, m)$$

Thus, if $m = n = 5$, the probability of only 2 runs is given by putting $k = 1$ in (13.5) and is

$$P\{u = 2\} = 2[C(4, 0)]^2 / C(10, 5) = 1/126$$

and the probability that the number of runs is equal to or less than 4 is

$$\begin{aligned} P\{u \leq 4\} &= (f_2 + f_3 + f_4) / C(10, 5) \\ &= \{2[C(4, 0)]^2 + 2C(4, 1)C(4, 0) + 2[C(4, 1)]^2\} / C(10, 5) \\ &= (2 + 8 + 32) / 252 = \frac{1}{6} \end{aligned}$$

For $m > 10$, u is approximately normal, with mean $1 + 2mn/(m+n)$ and variance $\frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}$.

Tables giving $P\{u \leq u'\}$ for $m \leq n \leq 20$ have been prepared by F. S. Swed and C. Eisenhart (Reference 1). They have also calculated a set of confidence limits for u . Since u is necessarily an integer, the probability cannot, in general, be adjusted exactly to a predetermined value. For the upper 95% limit, they give the *smallest* u' for which $P\{u \leq u'\} \geq 0.975$ and for the lower limit the *largest* u' for which $P\{u \leq u'\} \leq 0.025$. The probability for a value between these limits is approximately 0.95.

Table 47 is a slightly modified extract from these tables for the case $m = n$. This table may be used to assess the significance of observed runs above or below the central line on a control chart, or above and below the median value in a sequence of measurements. The hypothesis of randomness is rejected if the observed u is less than the lower limit, or greater than the upper limit.

TABLE 47. CONFIDENCE LIMITS FOR NUMBER OF RUNS IN SEQUENCE OF m 0's AND m 1's

m	90% limits		95% limits	
	<i>lower</i>	<i>upper</i>	<i>lower</i>	<i>upper</i>
5	4	8	3	9
6	4	10	4	10
7	5	11	4	12
8	6	12	5	13
9	7	13	6	14
10	7	15	7	15
11	8	16	8	16
12	9	17	8	18
13	10	18	9	19
14	11	19	10	20
15	12	20	11	21
16	12	22	12	22
17	13	23	12	24
18	14	24	13	25
19	15	25	14	26
20	16	26	15	27
21	17	27	16	28
22	18	28	17	29
23	18	30	17	31
24	19	31	18	32
25	20	32	19	33
26	21	33	20	34
27	22	34	21	35
28	23	35	22	36
29	24	36	23	37
30	25	37	23	39
31	26	38	24	40
32	26	40	25	41
33	27	41	26	42
34	28	42	27	43
35	29	43	28	44
36	30	44	29	45
37	31	45	30	46
38	32	46	31	47
39	33	47	31	49
40	34	48	32	50

For $m > 10$, the number of runs is approximately normally distributed, with mean $m + 1$ and variance $m(m - 1)/(2m - 1)$.

Example 1. If the 30 observations in Fig. 48, §12.17, are classified as *above* or *below* the central value (a or b), we obtain the sequence

b a b b b b a a b b a a a a b b b a a b a a b a a b b b b b,

for which $m = 13$, $n = 17$, $u = 13$. In this case m is not the same as n , and from the original tables we find that the 95% limits for u are 10 and 21. There is no need to worry about a lack of randomness. Limits 11 and 21 are given in Table 47 for $m = 15$, and as a general rule if m and n are reasonably large and not very different we can, as an approximation, use an average value for m in Table 47.

Example 2. A student opened a set of mathematical tables, with the entries blocked off in sets of five, and, starting anywhere, added the five terminal digits in each block of five numbers. The sums so obtained for 50 blocks were:

12, 15, 18, 30, 33, 25, 28, 22, 23, 17
 25, 18, 22, 13, 17, 18, 22, 25, 27, 30
 28, 32, 24, 27, 20, 22, 15, 18, 20, 23
 12, 15, 27, 30, 33, 25, 28, 20, 23, 17
 25, 18, 20, 13, 17, 18, 22, 33, 27, 30

Test this sequence for randomness.

The median of these numbers is 22, and if we assign a 's and b 's according as the numbers are above or below the median we get

b b b a a a a — a b a b — b b b — a a a a a a b — b b b a b b
 a a a a b a b a b b b b b — a a a

There are 21 b 's, 24 a 's and 5 numbers which lie on the median. We can fill in these five with four b 's and one a in any way we like, preferably (so as to be conservative in rejecting the hypothesis of randomness) increasing the number of runs. Thus, if we complete the set of a 's and b 's as below, we have $m = n = 25$, $u = 20$.

b b b a a a a b a b a b a b b b b a a a a a a b
 b b b b a b b a a a a b a b a b b b b b a a a

The number of runs is less than we should expect, but well within the 95% confidence limits, so that the hypothesis of randomness is not rejected.

By filling the first blank space with an a and the others with b 's, we get only 16 runs, which would mean rejection of the hypothesis of randomness at the 95% level. The reason for filling in the spaces so as to increase the number of runs is that, in most practical cases where we want to test randomness (as in quality control work), the deviation from expectation is more likely to be in the direction of fewer and longer runs than in that of more and shorter runs. Wear on a machine tool, for example, may cause a gradual increase in the diameter of a machine part, or a slight slip in the setting may cause a sudden increase. Both causes will tend to produce long runs. If therefore we make the number of runs as large as possible, we reduce the risk of having to look for trouble where really none exists.

13.8 Run Test of Difference Between Two Samples. Given two samples, each consisting of m values of a variate x , we can make a rough test of the hypothesis that they come from the same population. The method is to arrange the $2m$ values in order of magnitude and test whether this order is random in respect of the two samples. Since, if the two samples are significantly different, the number of runs will be smaller than if they come from the same population, a one-sided test is indicated. The lower 90% limit in Table 47 will therefore give a significance level of 5% and the lower 95% limit a significance level of $2\frac{1}{2}\%$.

For instance, the following data given by Snedecor (Reference 2) refer to daily gains in weight (lb) of two lots of calves, each lot on a different ration:

I 1.95, 2.17, 2.06, 2.11, 2.24, 2.52, 2.04, 1.95

II 1.82, 1.85, 1.87, 1.74, 2.04, 1.78, 1.76, 1.86

Placed in order of magnitude, with those from lot II underlined, these values are:

1.74, 1.76, 1.78, 1.82, 1.85, 1.86, 1.87, 1.95, 1.95, 2.04,
2.04, 2.06, 2.11, 2.17, 2.24, 2.52

The number of runs (of underlined or non-underlined values) is 4. From Table 47 the 5% significance level corresponds to $u = 6$, so that there is a significant lack of randomness at this level. In fact, the probability of 4 runs or fewer may be computed from equation (13.7) as $19/2145 = 0.009$, so that this number is actually significant at the 1% level.

The advantage of this test is that it assumes very little about the nature of the parent population — merely that the variate is continuous and that the samples are random and independent. The disadvantage is a lack of power, a very marked difference between the samples being required, as a rule, to produce a significant departure from randomness.

In the example given above there were two identical values (each 2.04), one from each sample. In whichever order these are placed, the number of runs is still 4. Sometimes, however, the number of runs will be affected by the placing of such identical pairs, and then all the possible orders should be tested.

13.9 Random Numbers. The choosing of a truly random sample, even from an artificial population of discs, balls, cards, or the like, is not easily accomplished, since the common methods of mixing and shuffling are often inadequate. When it is necessary to pick random samples from a crop in the ground, or a field subdivided into small plots, or a group of experimental animals, the task is much harder. If one relies on personal judgments, it is difficult to avoid a tendency to pick what seem to be *typical* rather than *random* samples. Experience has shown that it is advisable to use a set of numbers which have been thoroughly tested for randomness, and to rely on these numbers for picking the sample. Tables of random numbers have been compiled by Tippett, by Kendall and Babington Smith, and by Fisher and Yates (see Reference 3). A 4-page extract from the tables of Kendall and Babington Smith is given in the Appendix (Table V).

These tables consist of four-digit numbers, obtained by a mechanical process and tested in several ways for randomness. In order to pick out a random sample from a given finite population, the items of the population are numbered consecutively and blocks of consecutive 4-digit numbers are allotted to

each item, the size of the blocks being such that the whole range of 10,000 numbers is fairly well covered. Thus, if we want to pick a sample of 20 from a population of 300, we can allot blocks of 30 numbers to each item. To the first item will correspond numbers 0000 to 0029, to the second item, numbers 0030 to 0059, and so on, until to the 300th item correspond numbers 8970 to 8999. We then read off 20 consecutive numbers, starting anywhere in the table of random numbers (and disregarding numbers beginning with 9). Each number will fall in some block and will therefore represent some item in the population and that item is picked for the sample.

With a population of 100 or less we can read the random numbers as 2-digit numbers, each 4-digit number being split in two.

A similar procedure will enable us to choose samples from a continuous distribution grouped in classes. For example, suppose we wish to pick a random subsample of 50 Glasgow schoolgirls from the sample of 1000 described in Table 27, §7.7. We must divide the 10,000 possible random numbers into blocks, of sizes proportional to the frequencies in the various classes, as shown in Table 48, where the first block contains 10 numbers, the next 140, and so on. We then read off a list of 50 random numbers, such as 9327, 6908, 2511, 8268, 3768, 6735, 9214, 0740, . . . and for each number take any girl from the corresponding class — in this example, from classes 8, 6, 5, 7, 5, 6, 8, 4 . . .

TABLE 48. ALLOCATION OF RANDOM NUMBERS FOR SUBSAMPLING FROM
SAMPLE OF TABLE 27

<i>Class Mark (lb)</i>	<i>f</i>	<i>Block of Numbers</i>
(1) 29.5	1	0000-0009
(2) 33.5	14	0010-0149
(3) 37.5	56	0150-0709
(4) 41.5	172	0710-2429
(5) 45.5	245	2430-4879
(6) 49.5	263	4880-7509
(7) 53.5	156	7510-9069
(8) 57.5	67	9070-9739
(9) 61.5	23	9740-9969
(10) 65.5	3	9970-9999
	1000	

For use in sampling experiments we can make up a similar table corresponding to any given population. If the population distribution is continuous (normal, for example), we can use a table of areas to give the sizes of blocks of random numbers corresponding to specified intervals of the variate x . In so doing we are, of course, replacing our continuous distribution by a discontinuous one, but if there are 20 to 30 classes corresponding to the effective range of the distribution, the resulting approximation is quite satisfactory.

For example, if we want a normal distribution with $\mu = 20$ and $\sigma = 4$, we can choose the x -interval as unity, divide the whole range into intervals such as 6.5 to 7.5, 7.5 to 8.5, etc., write $z = (x - 20)/4$, and find the corresponding areas ΔA . These areas, multiplied by 10,000, will give the sizes of the corresponding blocks of 4-digit random numbers.

Tables of random normal numbers have been compiled. One such table is given in Dixon and Massey's "Statistical Analysis" (Reference 5 of §0.4). These tables can be used directly to give random samples from the particular normal population specified at the head of the table.

13.10 The Sign Test for Differences in Paired Samples. The ordinary t -test for the reality of an observed effect in paired samples assumes that all the pairs may be regarded as random samples from the same population of pairs. Sometimes this assumption cannot be made, the pairs having been observed under widely different conditions. A non-parametric test, based only on the *signs* of the differences, can be used in such cases, although naturally it is not as powerful a test as the t -test, since it uses less information and makes fewer assumptions. It is, however, very simple and easy to apply.

Suppose A and B are two materials or varieties or treatments to be compared, and let x_1 be a measurement on A and x_2 the corresponding measurement on B in the same pair. Let there be N differences $d_i = x_{1i} - x_{2i}$, and let r be the number of times the less frequent sign occurs in the set of d_i , so that $r \leq N/2$. It may happen that some differences are exactly zero — these are excluded and the sample size correspondingly reduced. Then the distribution of r is binomial with $\theta = \frac{1}{2}$, and the critical values of r , corresponding to assigned significance levels, can be obtained from tables of the binomial distribution. The null hypothesis here is that each d_i has a distribution with median 0 (this distribution not being necessarily the same for all the d_i). The hypothesis is rejected if r differs significantly from $N - r$.

In Table 40, §12.12, data were given on hemoglobin in anemic rats, and the t -test was used to test the significance of the observed effect of a change in diet. If the experiments on the various pairs had been carried out in different places, with different-sized rats, of different racial strains, and so on, the t -test would have been inapplicable. According to the sign test, $r = 4$, $N = 12$. If we suppose that the difference in diet could only *increase* x , if it has any effect at all, we are interested in the probability that the number of *negative* signs is 4 or less. This is given by

$$\sum_{r=0}^4 C(N, r) \left(\frac{1}{2}\right)^N = 0.194$$

and therefore the observed effect is, by this test, non-significant, as in fact it also is by the t -test. If we do not know which way the effect will go, we want the probability that r is 4 or less for *either positive or negative* signs, and this is double the previous probability.

Table 49 (See Reference 7) gives for various values of N approximate significance levels of r (for a two-tailed test). Values equal to or less than those given in the table are significant at the indicated level. For a one-tailed test, the significance level should be halved. For significance at the 5% level in the example quoted above, the number of minus signs would have to be 2 or less.

TABLE 49. CRITICAL VALUES OF r FOR THE SIGN TEST (TWO-TAILED)

N	5%	10%	N	5%	10%	N	5%	10%
9	1	1	36	11	12	63	23	24
10	1	1	37	12	13	64	23	24
11	1	2	38	12	13	65	24	25
12	1	2	39	12	13	66	24	25
13	2	3	40	13	14	67	25	26
14	2	3	41	13	14	68	25	26
15	3	3	42	14	15	69	25	27
16	3	4	43	14	15	70	26	27
17	4	4	44	15	16	71	26	28
18	4	5	45	15	16	72	27	28
19	4	5	46	15	16	73	27	28
20	5	5	47	16	17	74	28	29
21	5	6	48	16	17	75	28	29
22	5	6	49	17	18	76	28	30
23	6	7	50	17	18	77	29	30
24	6	7	51	18	19	78	29	31
25	7	7	52	18	19	79	30	31
26	7	8	53	18	20	80	30	32
27	7	8	54	19	20	81	31	32
28	8	9	55	19	20	82	31	33
29	8	9	56	20	21	83	32	33
30	9	10	57	20	21	84	32	33
31	9	10	58	21	22	85	32	34
32	9	10	59	21	22	86	33	34
33	10	11	60	21	23	87	33	35
34	10	11	61	22	23	88	34	35
35	11	12	62	22	24	89	34	36

For $N \geq 90$, r is approximately the integer next below $(N - 1)/2 - k(N + 1)^{1/2}$, with $k = 0.9800$ and 0.8224 for the 5% and 10% values, respectively.

The sign test can be extended somewhat to answer such questions as the following:

Is A better than B by $p\%$ or by u units? To answer these we merely increase every measurement (x_2) on B by $p\%$ or by u units, and compare the resulting set with the original measurements (x_1) on A , using a one-tailed test.

Example 3. Suppose the data in Table 50 represent yields in bushels of two varieties of apples, these varieties being grown on adjacent trees. Variety A is clearly superior (all the signs of $x_1 - x_2$ are positive). Is it better than variety B by 5 bushels?

The numbers in column 4 (excluding zeros) give $r = 2$, $N = 11$, and this value of r is significant at the 5% level. If we ask whether B is 6 bushels better, we get the numbers in column 5, giving $r = 5$, $N = 11$, which is non-significant. We are therefore prepared to say that A is better than B by as much as 5 bushels, but not by as much as 6 bushels.

TABLE 50. COMPARISON OF YIELDS OF TWO VARIETIES BY SIGN TEST

$x_1(A)$	$x_2(B)$	$x_1 - x_2$	$x_1 - (x_2 + 5)$	$x_1 - (x_2 + 6)$
13	11	2	-3	-4
12	6	6	1	0
10	3	7	2	1
6	1	5	0	-1
13	7	6	1	0
15	10	5	0	-1
19	9	10	5	4
10	4	6	1	0
11	3	8	3	2
11	6	5	0	-1
13	8	5	0	-1
9	5	4	-1	-2
14	7	7	2	1
12	6	6	1	0
12	4	8	3	2

13.11 Inequalities of the Tchebycheff Type. It was proved by Tchebycheff (and independently by Bienaymé) that for any population, no matter how queer the distribution (provided only that the variance σ^2 is finite), the probability that a random value of the variate x will differ from its expected value by as much as λ is not more than σ^2/λ^2 .

In symbols,

$$(13.8) \quad \Pr\{|x - \mu| \geq \lambda\} \leq \sigma^2/\lambda^2$$

For example, the probability of a deviation from the expected value of at least 3σ is never more than $\frac{1}{9}$. Of course, if we know the form of the distribution, we may be able to make this inequality much sharper. If the distribution is normal, the probability of a deviation numerically as great as 3σ is only 0.0027, but the point about the Tchebycheff inequality is that it is true (with the restriction about the variance) for *any* distribution.

If we apply this inequality to the variate $\bar{x} = \frac{1}{N} \sum_i x_i$, where the x_i are independent variates with the same distribution, the variance of \bar{x} is σ^2/N ,

and the inequality becomes

$$(13.9) \quad \Pr\{|\bar{x} - \mu| \geq \lambda\} \leq \sigma^2/N\lambda^2$$

Since any probability ≤ 1 , the right-hand side must be less than 1 to give a useful result. If we put it equal to $\frac{1}{2}$, we see that $\lambda = \left(\frac{2\sigma^2}{N}\right)^{1/2}$, so that $|\bar{x} - \mu|$ is of the order of $1/\sqrt{N}$. This means that the discrepancy between the sample mean and the population mean is of the order of unity divided by the square root of the sample size, a result which emphasizes the importance of using large samples to get accurate estimates.

Various other inequalities of a similar nature are known. For instance, if we know that the distribution has a single mode at μ_0 , we have the Gauss inequality

$$(13.10) \quad \Pr\{|x - \mu_0| \geq \lambda\tau\} \leq 4/(9\lambda^2)$$

where

$$\tau^2 = \sigma^2 + (\mu - \mu_0)^2$$

If the distribution has a fourth moment μ_4 , then, as shown by Robbins,

$$(13.11) \quad \Pr\{|\bar{x} - \mu| \geq \lambda\} \leq [\mu_4 + 3(N-1)\sigma^4]/N^3\lambda^4$$

Where the range of the variate is known to be bounded, upper limits can be given for the moments. For example, if a distribution of heights is known to lie between 64 and 78 in., the largest possible value of σ^2 is when half the population has height 64 in. and half 78 in., and is therefore $7^2 = 49$ in.². The corresponding value of μ_4 is $7^4 = 2401$ in.⁴. The probability that the mean height for a sample of 200 individuals from this population will differ from the mean of the population by as much as 1 in. is, from (13.9), not more than 0.25 ($\lambda = 1$, $N = 200$, $\sigma^2 = 49$), and, from (13.11), not more than 0.18.

13.12 Order Statistics. The statistics most commonly used, such as the mean and standard deviation, depend on linear or quadratic combinations of the various observed values without regard to their order. Other statistics, including the median and other percentiles and the range, depend on the ordering of the observed values according to magnitude. These statistics are spoken of as *order statistics* or *systematic statistics*.

Although much is known about the exact distributions of statistics of the first type, particularly for samples from a normal parent population, the exact distributions of order statistics are usually very troublesome to calculate. Tables have been calculated for a normal parent population, and some results are known for rectangular and other special populations, and for very small samples.

Reference 8 (at the end of the chapter) gives an interesting treatment of

some of the more important results in the sampling theory of order statistics and their applications to statistical inference.

13.13 The Median. If the parent population is symmetrical, the sample median may be regarded as an estimate of the population mean μ . It is an unbiased estimate, since its expected value is μ , but its sampling variance is often larger than the sampling variance of the arithmetic mean, and therefore it is not as efficient an estimate. For a normal population, the mean is most efficient (no other unbiased statistic for estimating μ can have a smaller variance), and the efficiency of the median is measured by the ratio of the variance of the mean to the variance of the median. This efficiency depends on the sample size, and for large N it tends to the value $2/\pi = 0.637$. That is to say, we can get, on the average, as good a value of μ from the mean of 637 observations from a normal population as from the median of 1000.

For small samples, the median is relatively more efficient. Thus for samples of size N , the efficiency is given by E in the following table:

N	3	4	5	6	7	8	9	10
E	0.74	0.84	0.69	0.78	0.67	0.74	0.65	0.71

The efficiency is higher when N is even because then the median is defined as the arithmetic mean of the two middle values.

Confidence limits for the median of the parent population obtained from a small sample have been calculated by K. R. Nair (Reference 4). These limits are non-parametric; they make no assumption about the nature of the parent population, except that the variate is continuous. We suppose the observations arranged in ascending order of magnitude, $x_1 < x_2 < \cdots < x_N$, and make a statement that the median lies between x_k and x_{N+1-k} . The 95% confidence limits are given by the largest k for which the probability that the statement is true is at least 0.95. Thus, if $N = 20$, we find $k = 6$. The probability, *before* drawing a sample of 20, that the median of the population will lie between the 6th and the 15th observations, when these are arranged in order, is at least 0.95 and, in fact, is 0.959. An extract from Nair's table is given in Table 51.

The distribution of the median of samples from a population with a given frequency function $f(x)$ can be calculated, under certain assumptions, for the limit when the size of the sample tends to infinity. Let the number of elements in the sample be $2n + 1$, so that the median is the value x_{n+1} . If \tilde{x}_0 is the population median, so that

$$\int_{-\infty}^{\tilde{x}_0} f(x) dx = 1/2$$

then it can be proved that for large values of n the sample median \tilde{x} is approximately normally distributed with mean \tilde{x}_0 and variance $1/[8nf^2(\tilde{x}_0)]$ where

TABLE 51. APPROXIMATE 95% CONFIDENCE INTERVALS FOR THE MEDIAN
(The confidence coefficient is P that the population median lies between x_k and x_{N-k+1})

N	k	P	N	k	P
10	2	0.979	31	10	0.971
11	2	.988	32	10	.980
12	3	.961	33	11	.965
13	3	.978	34	11	.976
14	3	.987	35	12	.959
15	4	.965	36	12	.971
16	4	.979	37	13	.953
17	5	.951	38	13	.966
18	5	.969	39	13	.976
19	5	.981	40	14	.962
20	6	.959	41	14	.972
21	6	.973	42	15	.956
22	6	.983	43	15	.968
23	7	.965	44	16	.951
24	7	.977	45	16	.964
25	8	.957	46	16	.974
26	8	.971	47	17	.960
27	8	.981	48	17	.971
28	9	.964	49	18	.956
29	9	.976	50	18	.967
30	10	.957			

$f^2(\tilde{x}_0)$ is the square of $f(\tilde{x}_0)$. This is called an *asymptotic distribution*. If the parent population is normal, then

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

and $\tilde{x}_0 = \mu$. Hence $f(\tilde{x}_0) = 1/[\sigma\sqrt{2\pi}]$, so that the variance of \tilde{x} is

$$2\pi\sigma^2/8n = \pi\sigma^2/4n$$

Since the variance of \tilde{x} is $\sigma^2/(2n+1)$, the efficiency of the median is

$$4n/[\pi(2n+1)]$$

or approximately $2/\pi$.

13.14 Estimation by Percentiles. The average of the quartiles $(Q_1 + Q_3)/2$, or $(P_{25} + P_{75})/2$ in the notation of percentiles, is an efficient estimate of the mean of a normal distribution, the efficiency being around 0.87 for samples of 10 and falling off slowly to 0.81 for very large samples. Still higher efficiencies may be obtained by using more percentiles, for instance, 0.88 for $(P_{17} + P_{50} + P_{83})/3$ in large samples. These and other percentiles may be very rapidly found mechanically when the data are entered on cards, one card for each value. The cards are arranged in order of magnitude of

the observations and run through a sorting machine. If we have 300 cards, we run 51 through the machine and read the 51st card, run another 99 through and read the 150th, and finally run yet another 99 through and read the 249th. These three cards will give P_{17} , P_{80} , and P_{83} respectively.

Dispersion also may be estimated by percentiles. If only two are used, the sampling variance of the estimate of σ is least when the percentiles used are P_{07} and P_{93} , and is then 0.65 for large samples. For a normal population, the interval $P_{93} - P_{07}$ is equal to 2.952σ , as is easily found by interpolation in the table of areas for the standard normal curve. The estimate of σ from these two percentiles is $(P_{93} - P_{07})/2.952 = 0.3388(P_{93} - P_{07})$. If four percentiles are used, the best estimate of σ is $0.1714 (P_{97} + P_{85} - P_{15} - P_{03})$, with an efficiency of 0.80 .

If we want to estimate both the mean and the standard deviation by the same percentiles, we have to compromise between the requirements for these two statistics separately. The percentiles which give the best estimates of μ are not those which give the best estimates of σ . For estimation with *two* percentiles, we can take

$$(13.12) \quad \begin{cases} \hat{\mu} = (P_{15} + P_{85})/2, \text{ efficiency } 0.73 \\ \hat{\sigma} = 0.4824 (P_{85} - P_{15}), \text{ efficiency } 0.56 \end{cases}$$

and for estimation with *four* percentiles,

$$(13.13) \quad \begin{cases} \hat{\mu} = (P_{05} + P_{30} + P_{70} + P_{95})/4, \text{ efficiency } 0.80 \\ \hat{\sigma} = 0.2305 (P_{95} + P_{70} - P_{30} - P_{05}), \text{ efficiency } 0.74 \end{cases}$$

Percentile estimation is particularly useful in situations where it is comparatively easy to arrange the sample values in order, but much more troublesome to carry out the actual measurements. The exact measurements are required only for the percentile values used in the estimation, and not for the whole sample.

13.15 The Range. The range is the difference between the largest and smallest observations in the sample, namely

$$(13.14) \quad R = x_N - x_1$$

As an estimate of dispersion for samples from a normal population the range is very poor for large N , but for small samples its efficiency is remarkably high. This, of course, does not mean that the range of a very small sample will give a very good estimate of the population standard deviation, it means merely that the range is almost as good in this respect as the sample standard deviation. With a small sample *no* estimate can be very precise. Table 52 gives the values of the multiplier k used in estimating σ from R , and also the corresponding efficiencies, for values of N from 2 to 10. The range is seldom used for large samples.

TABLE 52. ESTIMATION FROM THE RANGE FOR A NORMAL PARENT POPULATION

<i>N</i>	<i>k</i>	<i>E</i>	99% Confidence Limits for <i>R</i> / <i>σ</i>	
			lower	upper
2	0.886	1.00	0.01	3.97
3	.591	0.99	.13	4.42
4	.486	0.98	.34	4.69
5	.430	0.96	.55	4.89
6	.395	0.93	.75	5.03
7	.370	0.91	.92	5.15
8	.351	0.89	1.08	5.26
9	.337	0.87	1.21	5.34
10	.325	0.85	1.33	5.42

$$\hat{\sigma} = kR, \text{ efficiency} = E$$

This table may be used in forming limits for a control chart, the mean range *R* being obtained from a considerable number of samples each of size *N*, and the standard deviation of the population estimated from

$$(13.15) \quad \hat{\sigma}_1 = k\bar{R}$$

The distribution function for the range in samples from a population with known frequency function is expressible in terms of integrals but is difficult to evaluate, even for a normal population. The only important case for which the range turns out to have a simple distribution function is that of a *rectangular* parent population (one for which $f(x) = 1/c$, $0 < x < c$, and $f(x) = 0$ for all other values of x). The distribution function for the range *R* of samples from a rectangular population, that is, the probability for a value less than *R*, is given by

$$(13.16) \quad F(R) = N \left(\frac{R}{c} \right)^{N-1} - (N-1) \left(\frac{R}{c} \right)^N$$

The probability of a value between *R* and *R* + *dR* is

$$(13.17) \quad f(R)dR = N \frac{(N-1)}{c} \left(\frac{R}{c} \right)^{N-2} \left(1 - \frac{R}{c} \right) dR$$

For the *normal* law, H. O. Hartley (Reference 5) has calculated tables giving the probability $F(R)$ for values of *N* between 2 and 20. Upper and lower 99% confidence limits for the ratio of *R* to *σ* are given in Table 52 for sample sizes from 2 to 10.

13.16 Quotient of Ranges in Samples from a Rectangular Population. A rectangular population is not quite as artificial as it appears at first sight. It has been asserted that errors of observation in accurate physical measure-

ments are in reality more nearly rectangular than normal — very large errors are not merely rare but do not occur at all (apart from gross mistakes, which are usually obvious). Also, in the production of machine parts in a factory to rather narrow specification limits, when only those articles which comply with the specification are included in the population, it appears that the hypothesis of a rectangular distribution is not unreasonable.

Let us suppose that we have two random samples of sizes m and n , with ranges R_1 and R_2 , from the population specified above. The distribution of $R_1/R_2 = u$ was worked out by Rider (Reference 6) and turns out to be independent of c . The frequency curve is a skew curve, with the mode at

$$(13.18) \quad u = (m-2)(m+n)/(m-1)(m+n-2), \quad m-n \leq 2$$

or

$$(13.19) \quad u = (n+1)(m+n-2)/n(m+n), \quad m-n \geq 2$$

and the mean at

$$(13.20) \quad \mu_1' = (m-1)n/(m+1)(n-2)$$

Its equation is

$$(13.21) \quad f(u) = \frac{m(m-1)n(n-1)}{(m+n)(m+n-1)(m+n-2)} \times \\ [(m+n)u^{m-2} - (m+n-2)u^{m-1}]$$

for $0 \leq u \leq 1$, and

$$(13.22) \quad f(u) = \frac{m(m-1)n(n-1)}{(m+n)(m+n-1)(m+n-2)} \times \\ [(m+n)u^{-n} - (m+n-2)u^{-n-1}]$$

for $1 \leq u < \infty$.

Table 53, giving values for the quotient of ranges which will be exceeded in 5% of random samples, may be used to test the hypothesis that two samples come from the same rectangular population.

Example 4. The width of a slot in a certain airplane part was measured to the thousandth of an inch in a sample of 5 parts on the first day of production, and again in a sample of 10 two days later. The results (in thousandths of an inch in excess of 0.800 in.) were

I 77, 80, 78, 72, 78, ($R_1 = 8$)

II 75, 77, 75, 76, 77, 79, 75, 78, 77, 76, ($R_2 = 4$)

The null hypothesis is that both samples come from the same rectangular population. Taking m as the number in the sample with the larger range, we have

$$u = 2, \quad m = 5, \quad n = 10$$

TABLE 53. 5% POINTS FOR THE DISTRIBUTION OF THE QUOTIENT OF RANGES FROM A RECTANGULAR POPULATION

<i>n</i>	<i>m</i> = number in sample with greater range							
	3	4	5	6	7	8	9	10
3	4.00	4.64	5.08	5.39	5.63	5.82	5.97	6.09
4	2.31	2.62	2.83	2.99	3.11	3.20	3.28	3.34
5	1.75	1.96	2.10	2.20	2.28	2.34	2.40	2.44
6	1.49	1.64	1.75	1.83	1.89	1.94	1.98	2.01
7	1.34	1.46	1.55	1.61	1.66	1.70	1.73	1.76
8	1.24	1.35	1.42	1.47	1.52	1.55	1.58	1.60
9	1.17	1.27	1.33	1.38	1.42	1.45	1.47	1.49
10	1.13	1.21	1.27	1.31	1.34	1.37	1.39	1.41

From Table 53, the 5% point is 1.27, so that the value of u is significant. The actual probability of getting a value of u as large as 2 is 0.0013, which can be calculated by integrating equation (13.22) for the distribution of u . The conclusion is that the second sample is significantly more uniform than the first. The machine has settled down to a steadier production.

Note that this test is analogous to the F -test discussed in §12.14. The null hypothesis is that the quotient of ranges is 1, and it is tested against the alternative hypothesis that the quotient is greater than 1. If we want to test the null hypothesis against the alternative that the quotient is either greater than or less than 1, the 5% level becomes a 10% level. In Example 4, we should reject the null hypothesis at the 5% level if either

$$u > 1.27 (m = 5, n = 10)$$

or

$$1/u > 2.44 (m = 10, n = 5)$$

and the probability of the combined event is 0.10.

Exercises

1. Test the goodness of fit of the Poisson distribution to results of a sampling experiment in Table 36, §10.8. Note that two sets of calculated frequencies are given, one corresponding to an assumed value of λ and the other to an estimated value. The degrees of freedom are one fewer in the latter case.

2. Test the goodness of fit of the Poisson distributions fitted in Exercises 16 and 17, page 160.

3. The distribution of Table 25, page 87, was graduated by means of a normal curve in Exercise 16, page 122. Test the goodness of fit.

4. Try for yourself the procedure described in Example 2, §13.7, for forming a set of random numbers (all these numbers lie between 0 and 45, inclusive). Collect 80 such numbers and test the sequence for randomness.

5. A student dealt 26 cards from an ordinary deck 50 times and each time counted the number of Honors in the 26 cards dealt. (A, K, Q, J, 10 counted as Honors.) The distribution obtained was:

x	4	5	6	7	8	9	10	11	12	13	14	15	16	17
f_o	1	0	2	3	7	5	10	8	2	7	2	0	2	1

Would you reject the hypothesis that the cards were well shuffled between each deal?

Hint. The probability on this hypothesis of x honor cards in 26 is $C(20, x)C(32, 26 - x)/C(52, 26)$. If this probability is calculated and multiplied by 50, the theoretical distribution is

x	4	5	6	7	8	9	10	11	12	13	14	15	16
f_e	0.0	0.2	0.9	2.7	6.0	9.6	11.2	9.6	6.0	2.7	0.9	0.2	0.0

Compare the two distributions by the χ^2 test.

6. In the following table x is the number of fives or sixes observed in a single throw with five dice:

x	0	1	2	3	4	5
f_o	23	90	81	30	19	0

Would you reject the hypothesis that the dice are true?

Hint. The probability on the null hypothesis of x fives or sixes with 5 dice is $C(5, x) \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{5-x}$.

7. Two samples of machine parts gave the following measurements (unit thousandth of an inch):

I	801,	798,	800,	805,	800,	804
II	796,	797,	796,	798,	794,	795

Do these differ significantly by the run test?

Ans. Number of runs is 2 or 4. The 5% significance value is 3.

8. Use the table of random numbers (Appendix Table V) to select a set of 100 random samples of 10 from the population of Table 38, §12.2. Find the mean of these samples and compare with the mean of the parent population.

Hint. Form a table of blocks of random numbers like that in Table 48.

9. Use the table of random numbers and the blocks given in Table 48 to draw a subsample of 100 from the population of 1000 Glasgow schoolgirls. Form this subsample into a frequency distribution and calculate its mean and variance. Compare with the theoretical values.

10. The following distribution of the range was obtained in 200 samples of 10 from an artificially constructed, approximately normal population with mean 20 and standard deviation 4:

R	5	6	7	8	9	10	11	12	13
f	2	4	4	14	11	20	25	28	25

R	14	15	16	17	18	19	20	21	22	23
f	17	13	13	5	9	3	3	3	0	1

Calculate the mean range and estimate the standard deviation of the population. Compare with the true value 4.

Estimate the 99% limits for R/\bar{R} from this sample and compare with the values given in Table 44, §12.17, for $N = 10$.

11. A distribution is symmetrical with a single mode at $x = 0$ and is bounded between $x = -b$ and $x = b$. Find 95% limits for the deviation of the mean of a sample of 100 from the population mean.

Ans. $|\bar{x}| \leq 0.16b$, with a probability of 0.95.

Hint. Use inequality (13.11). The greatest possible values of σ^2 and μ_4 will be less than the values for a rectangular distribution $f(x) = 1/2b$, $-b < x < b$. For this distribution, $\sigma^2 = b^2/3$ and $\mu_4 = b^4/5$. Put the probability that $|\bar{x} - 0| \geq \lambda$ equal to 0.05 and solve the equation so obtained for λ .

References

1. F. S. Swed and C. Eisenhart, "Tables for Testing Randomness of Grouping in a Sequence of Alternatives," *Ann. Math. Stat.*, **14**, 1943, pp. 66-87.
2. G. W. Snedecor, *Statistical Methods*, 4th ed. (Iowa State College Press, 1946), p. 78.
3. L. H. C. Tippett, "Random Sampling Numbers." (*Tracts for Computers*, No. 15), Cambridge University Press, 1927.
- M. G. Kendall and B. Babington Smith, "Tables of Random Sampling Numbers." (*Tracts for Computers*, No. 24), Cambridge University Press, 1946.
- R. A. Fisher and F. Yates, *Statistical Tables for Biological Agriculture and Medical Research*, 3rd ed. (Oliver and Boyd, 1949), pp. 104-109.
4. K. R. Nair, "Table of Confidence Intervals for the Median in Samples from any Continuous Population," *Sankhyā*, **4**, 1940, pp. 551-558. This table is reproduced in Dixon and Massey's *Introduction to Statistical Analysis*, p. 360.
5. H. O. Hartley, "The Range in Random Samples," *Biometrika*, **32**, 1942, p. 309.
6. P. R. Rider, "Distribution of the Quotient of Ranges in Samples from a Rectangular Population," *J. Amer. Stat. Assoc.*, **46**, 1951, pp. 502-507.
7. W. J. Dixon and A. M. Mood, "The Statistical Sign Test," *J. Amer. Stat. Assoc.*, **41**, 1946, pp. 557-566.
8. S. S. Wilks, "Order Statistics," *Bull. Amer. Math. Soc.*, **54**, 1948, pp. 6-50.
9. For an excellent discussion of the theory and practical use of the Chi-square test, see W. J. Cochran, "The χ^2 Test of Goodness of Fit," *Ann. Math. Stat.*, **23**, 1952, pp. 315-345.

CHAPTER XIV

TIME SERIES

14.1 Time as a Variable. Hitherto we have considered the distribution of a single variable, first from the descriptive standpoint and then from the point of view of estimation and significance. We now have to take up problems in which there are two variable quantities, and in the present chapter one of the two variables will be time. A set of data depending on the time is called a *time series*.

The time variable, of course, does not fluctuate arbitrarily. It moves uniformly, always in the same direction, from past to future. We can often, however, exercise some freedom of choice as to the times at which we make observations, although in most instances it is convenient to observe at regular intervals.

In the typical time series there are discernible three main features which

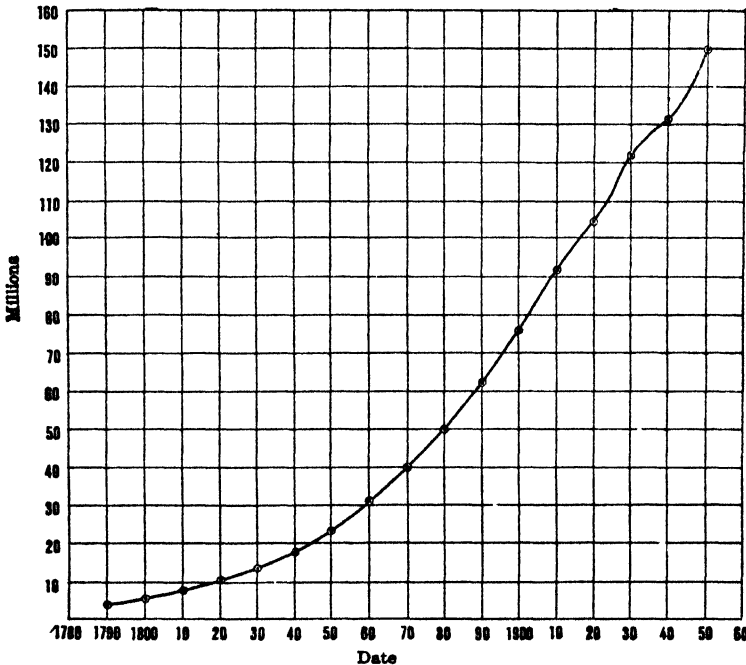


FIG. 49. POPULATION OF CONTINENTAL U.S.A.

seem to be independent of one another and attributable to distinct causes: (1) a broad long-term movement, called the *trend*, such as a more or less steady rise or fall; (2) an *oscillation* about the trend, which may be a seasonal effect with fairly regular period or a rather long-period, somewhat irregular oscillation, often called a *cycle*; (3) an irregular, unsystematic or random component, sometimes called the *residual*.

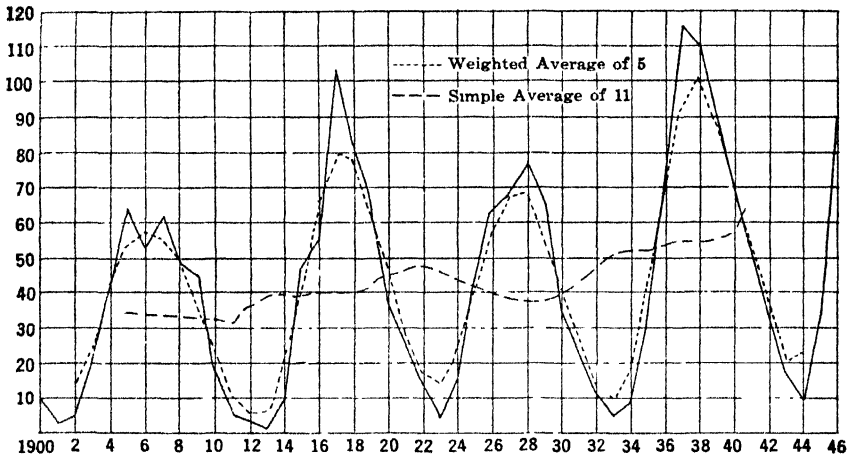


FIG. 50. RELATIVE SUNSPOT NUMBERS

Not all time series exhibit all three of these features. Three series are graphed in Figs. 49, 50 and 51. Fig. 49 gives the population of the U.S.A. at decennial censuses since 1790 and is mainly trend with a small residual element of randomness. Fig. 50 gives the mean annual sunspot number (Wolf and Wolfer system) for fifty years, and the cyclical aspect is much in evidence, combined with a random effect. There is very little evidence of trend. Fig. 51 gives the total annual precipitation at Edmonton, 1900–1950, and is evidently almost entirely random.

One task of time series analysis is to disengage these separate elements from a given set of data, so as to exhibit the trend and the oscillations, if any, apart from the random fluctuations. Trends and cycles are important from the point of view of interpolation and also of attempted forecasts. The trend of population statistics can be interpolated to give estimates of the population for years in which no census is taken. Estimates so made are automatically corrected by the next following census. Also, if an economic time series shows a well-marked trend, with or without a superimposed “business cycle,” it may be worth while to extrapolate for a year or two into the future. But the results should be interpreted very cautiously, as they imply an assumption about the continuance of the causes which have led to

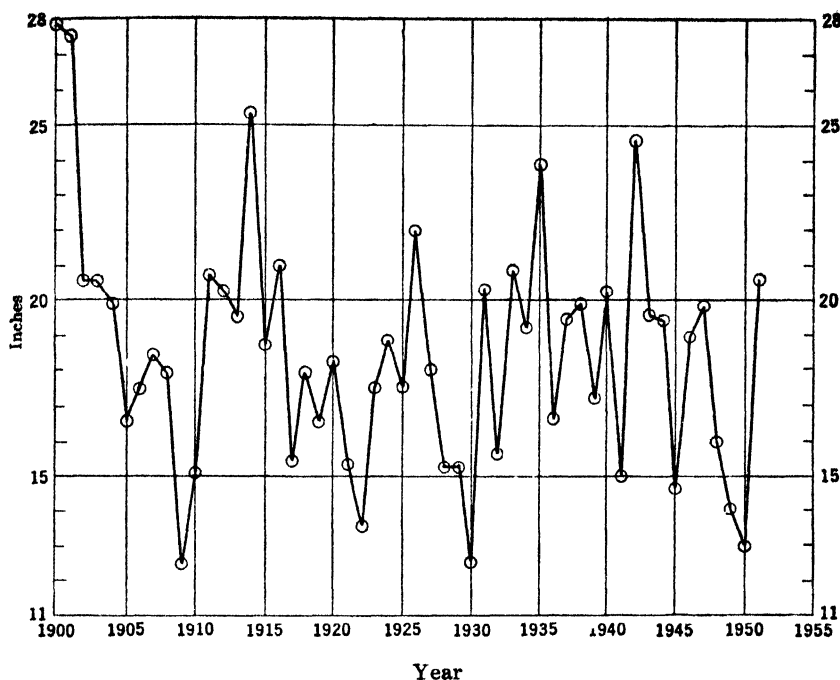


FIG. 51. TOTAL PRECIPITATION — EDMONTON, 1900-1951

the trend or the cycle, and, in the world of economics, conditions do not normally remain unchanged for very long.

14.2 Moving Averages. One method of smoothing out irregularities in a series in order to exhibit the trend is that of *moving averages*. (If there are pronounced seasonal fluctuations in the data, these should be removed first, in a way which will be described later.) We suppose for convenience that the successive observations are made at equal intervals of time which we will take as the unit of time (usually a year or a month), and we call the successive values y_0, y_1, y_2, \dots . If we need to go back in time before the instant we choose as our zero, we can write y_{-1}, y_{-2} , etc. For a *simple moving average* of 5 we take the mean of y_0, y_1, y_2, y_3 , and y_4 , and place it at $x = 2$ (the middle time), then the mean of y_1, y_2, y_3, y_4 , and y_5 and place it at $x = 3$, and so on all along our series. Calling these means $\bar{y}_2, \bar{y}_3, \dots$, we have

$$(14.1) \quad \begin{cases} \bar{y}_2 = (y_0 + y_1 + y_2 + y_3 + y_4)/5 = S_2/5 \\ \bar{y}_3 = (y_1 + y_2 + y_3 + y_4 + y_5)/5 = S_3/5 \\ \bar{y}_4 = (y_2 + y_3 + y_4 + y_5 + y_6)/5 = S_4/5 \end{cases}$$

In practice it is usually quicker, especially with a calculating machine, to

get each successive sum from the previous one by adding one new term and subtracting one old one. Thus,

$$S_3 = S_2 - y_0 + y_5$$

$$S_4 = S_3 - y_1 + y_6$$

...

Instead of a simple moving average, a *weighted* average is often used. The coefficients (or weights) are generally binomial. Thus, a weighted average of 5 would use the coefficients 1, 4, 6, 4, 1 which arise in the expansion of the binomial $(q + p)^4$. The formula for the successive terms of the sequence of averages is

$$(14.2) \quad \begin{cases} \bar{y}_2 = (y_0 + 4y_1 + 6y_2 + 4y_3 + y_4)/16 \\ \bar{y}_3 = (y_1 + 4y_2 + 6y_3 + 4y_4 + y_5)/16 \\ \dots \end{cases}$$

Such an average is most easily computed by taking a binomial average of 3 and another binomial average of 3 of the first set of averages.

If

$$\bar{y}_1 = (y_0 + 2y_1 + y_2)/4$$

$$\bar{y}_2 = (y_1 + 2y_2 + y_3)/4$$

$$\bar{y}_3 = (y_2 + 2y_3 + y_4)/4$$

and if

$$\bar{\bar{y}}_2 = (\bar{y}_1 + 2\bar{y}_2 + \bar{y}_3)/4$$

$$\bar{\bar{y}}_3 = (\bar{y}_2 + 2\bar{y}_3 + \bar{y}_4)/4$$

...

it is easily seen that $\bar{\bar{y}}_2 = (y_0 + 4y_1 + 6y_2 + 4y_3 + y_4)/16$, etc.

The use of the weighted rather than the simple average tends to produce a smoother curve while preserving the main features of the time series. An example of the computation for a weighted average of 5 for the sunspot data of Fig. 50 is given in Table 54, and the results are plotted in Fig. 50. The moving average has evidently smoothed out the random variations while preserving the general character of the oscillation.

How many terms should be included in the moving average is a matter of judgment in each particular case. If it is desired to smooth out an oscillation of regular period, the number of terms should exactly cover the period. The period of the sun-spot cycle is about 11.2 years, but is rather irregular. The effect of a simple moving average of 11 is shown in Fig. 50. For many commercial data (sales, etc.) there is a well-marked seasonal effect which is smoothed out by taking an average of 12 for monthly data. As a rule it is best to use an odd number of terms in a moving average so that the average can be attached to the time coordinate of the middle term. With the seasonal

TABLE 54. CALCULATION OF BINOMIALLY-WEIGHTED AVERAGE OF 5 FOR SUNSPOT DATA

Year (x_i)	Mean Sunspot Number (y_i)	Sum of 3 ($y_{i-1} + 2y_i + y_{i+1}$)	Average of 3 \bar{y}_i	Sum of 3 ($\bar{y}_{i-1} + 2\bar{y}_i + \bar{y}_{i+1}$)	Average of 3 $\bar{\bar{y}}_i$
1900	9.5				
01	2.7	19.9	5.0		
02	5.0	37.1	9.3	47.6	11.9
03	24.4	95.8	24.0	100.3	25.1
04	42.0	171.9	43.0	165.7	41.4
05	63.5	222.8	55.7	212.7	53.2
06	53.8	233.1	58.3	228.9	57.2
07	62.0	226.3	56.6	222.2	55.5
08	48.5	202.9	50.7	196.7	49.2
09	43.9	154.9	38.7	149.8	37.5
1910	18.6	86.8	21.7	90.5	22.6
11	5.7	33.6	8.4	42.1	10.5
12	3.6	14.3	3.6	19.6	4.9
13	1.4	16.0	4.0	28.6	7.1
14	9.6	68.0	17.0		
15	47.4				

data the moving average can be centered on a month by taking an average of 2 of an average of 12, which will attach the final average to the 7th month of the original 12.

More complicated moving averages are sometimes used by actuaries in smoothing long series. One such is Spencer's 15-point formula, with weights $-3, -6, -5, 3, 21, 46, 67, 74, 67, 46, 21, 3, -5, -6, -3$. This can be obtained by taking a weighted average of 5 with weights $-3, 3, 4, 3, -3$, then a simple average of 5, and finally two simple averages of 4.

One defect in a moving average of $2k + 1$ terms is that we lose k terms at the beginning and k at the end. This may be serious in comparatively short series and is an argument for using small values of k .

14.3 The Slutsky-Yule Effect. If a moving average is used to determine trend, it will also have an effect on the genuinely oscillatory component (if any) of the time series. A long-period oscillation tends to be included as part of the trend, whereas oscillations comparable in period with the length of the moving average or even shorter are damped out. But there is also an effect on the purely random component. It was proved by Slutsky and by Yule (independently) that a moving average may generate an irregular oscillatory movement where none existed in the original data. This is the Slutsky-Yule effect. For a simple moving average of length k , the variance of the induced oscillation is $1/k$ times the variance of the random component, and

the average length of the oscillation is $360/\theta$ where θ is the angle (in degrees) in the first quadrant with cosine $(k-1)/k$. The effect is increased when weighted averages are used. It is necessary in discussing the reality of possible oscillatory components in a time series to consider whether or not they may be spurious.

14.4 Mathematical Trend Lines. A trend obtained by the method of moving averages, even though fairly smooth, is not, as a rule, conveniently representable by a mathematical equation. For interpolation and extrapolation the advantages of a mathematically expressed trend line are obvious. The attempt is therefore often made to fit the observed data with a fairly simple curve, and the simplest of all is the straight line. Even when a straight line clearly will not fit over a long interval of time it is sometimes possible to break the interval up into subintervals over each of which the trend is approximately linear, and thus to "splice" two or more trend lines together. In other circumstances we may try to fit parabolic, cubic, or even higher polynomial curves, or use an exponential curve. Still more complicated curves are often used by actuaries in smoothing population and mortality statistics. We consider first the straight line trend.

14.5 Linear Functions. We know from algebra that the general form of a linear equation in two variables is

$$Ax + By = C$$

where A , B and C are arbitrary constants.

When $B \neq 0$, the equation may be solved for y , giving $y = -(A/B)x + C/B$ which is of the form

$$(14.3) \quad y = a + bx$$

and which is the form we will ordinarily use to represent a straight line.

The special cases where A or B or C is zero are as follows:

When $A = 0$, then $y = C/B$, which is of the form $y = a$. This is a line parallel to the x -axis. When $B = 0$, the equation takes the form $x = C/A$ which is a line parallel to the y -axis. When $C = 0$, then $Ax + By = 0$ which is a line passing through the origin.

The graph of (14.3) is a straight line (which explains the term "linear"). A characteristic property of a linear function is revealed at once by its graph. This is the fact that the ratio of a change in y to the corresponding change in x is *constant*. Thus, if two points (x_1, y_1) and (x_2, y_2) are chosen on the line, the value of the ratio

$$b = \frac{y_2 - y_1}{x_2 - x_1}$$

is independent of the points chosen. This ratio gives the average rate of

change of any function over the interval $\Delta x = x_2 - x_1$. In the case of a linear function, b defines the rate of change of the function.

Graphically, b is the slope of the line. If the units of x and y are identical and the scales are the same, b is the tangent of the angle of inclination θ which the line makes with the positive x -axis.* Lines having the same slope are parallel, and conversely.

It is shown in analytic geometry that we may obtain the slope of a straight line from its equation if we solve for y and take the coefficient of x . Thus in $2x - y = 5$, $y = 2x - 5$ and the slope is 2.

Conversely, if we know the slope of a line and the coordinates of any point on the line we can write its equation from the relation

$$(14.4) \quad y - y_1 = b(x - x_1)$$

which is called the *point-slope* form of a straight line. Thus, given that $(2, -1)$ is a point on a line whose slope is 2, the equation of the line is therefore $y + 1 = 2(x - 2)$ or $2x - y = 5$.

Or again, remembering that b is defined by a ratio involving the coordinates of two points on a line, we can obtain the equation of a line if we know any two points which lie on it. From the definition of b and (14.4), we have

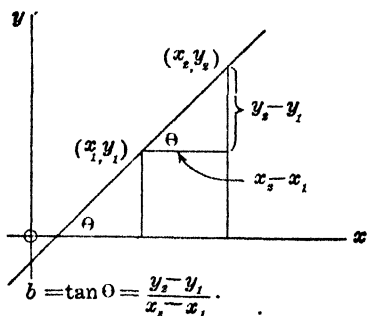
$$(14.5) \quad y - y_1 = \frac{y_2 - y_1}{x_2 - x_1} (x - x_1)$$

which is known as the *two-point* form of the equation. Thus, given that $(2, -1)$ and $(6, 7)$ are two points on a line, its equation is

$$y + 1 = \frac{7 + 1}{6 - 2} (x - 2) \quad \text{or} \quad 2x - y = 5$$

14.6 Fitting a Straight Line. The preceding discussion is intended as a basis for the presentation of certain methods of fitting a line to data. The equation $y = a + bx$ represents a family or set of lines corresponding to different values of the arbitrary constants a and b . The process of finding the best fitting line for any given data consists in determining a and b . By "best fitting" we mean best under a criterion of approximation specified by a method. We will consider three such methods: (a) *graphical*, (b) *the method of moments of ordinates*, (c) *the method of least squares*.

* When the line is vertical, $\theta = 90^\circ$ and b does not exist. Then $\Delta x = 0$ and division by zero is excluded in our algebra. With time series the units are usually different for x and y and the inclination has no physical meaning. The slope is the important quantity.

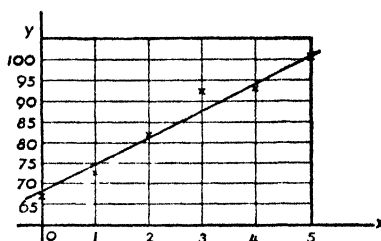


14.7 Graphical Method. A straight line is drawn (preferably with the aid of a transparent ruler) to fit as closely as possible the plotted points. To find the equation of this line, select two points on the line and estimate their coordinates (x_1, y_1) and (x_2, y_2) . Substituting these coordinates in the "two-point" form of the line (14.5), we get the desired equation.

If the first point is chosen so that $x_1 = 0$ the numerical work of simplifying the equation is somewhat lessened.

Example 1. Fit a line graphically to the following data.

x		y
(1948)	0	66.7
	1	72.7
	2	82.3
	3	92.1
	4	93.0
(1953)	5	100.6



We take the origin of x at 1948, hence from the figure $(x_1 = 0, y_1 = 68)$ and $(x_2 = 5, y_2 = 101)$.

By equation (3),

$$y - 68 = \frac{101 - 68}{5} x$$

Therefore,

$$y = 6.6x + 68$$

is the required equation.

The graphical method is open to the objection that it depends upon the judgment of the investigator. Different people will locate the line in different positions and therefore obtain different equations. However, where only approximate results are needed it is usually quite satisfactory.

14.8 Method of Moments. The constants a and b in the equation of a straight line fitted mathematically to a given time series are statistics calculated from the data. They may be regarded as estimates of the parameters α and β of the "true" trend line

$$(14.6) \quad y = \alpha + \beta x$$

To avoid confusion we will use the symbol y_i for the *observed value* of the variate at time x_i , and Y_i for the calculated *trend value* given by

$$(14.7) \quad Y_i = a + bx_i$$

There are two common methods of fitting a mathematical trend line. In the method of moments, the constants of the line are chosen so that the y_i and the Y_i have the same zeroth and first moments about the origin of x ,

where the r th moment of y_i is defined by $\sum_i y_i x_i^r$. That is to say, for the straight line,

$$(14.8) \quad \sum_i y_i = \sum_i Y_i = \sum_i (a + bx_i)$$

and

$$(14.9) \quad \sum_i y_i x_i = \sum_i Y_i x_i = \sum_i (a + bx_i)x_i$$

If the number of observations is N , these equations may be written

$$(14.10) \quad \begin{cases} \sum y_i = Na + b \sum x_i \\ \sum y_i x_i = a \sum x_i + b \sum x_i^2 \end{cases} .$$

and they are called the *normal equations* of the problem. They are a pair of simultaneous linear equations for the unknowns a and b .

Example 2. Find by the method of moments the best fitting line for the data in Example 1.

x	y	xy	x^2
0	66.7	0	0
1	72.7	72.7	1
2	82.3	164.6	4
3	92.1	276.3	9
4	93.0	372.0	16
5	100.6	503.0	25
15	507.4	1388.6	55

The normal equations are

$$(14.11) \quad \begin{cases} 6a + 15b = 507.4 \\ 15a + 55b = 1388.6 \end{cases}$$

These may be solved by eliminating a and obtaining an equation for b . Thus, if we multiply the first equation by 5 and the second by 2, we get

$$30a + 75b = 2537.0$$

$$30a + 110b = 2777.2$$

Subtracting the first of these from the second, we obtain

$$35b = 240.2$$

whence $b = 6.86$. The first of (14.11) then gives $6a = 507.4 - 102.9 = 404.5$ so that $a = 67.4$. The line is therefore

$$Y = 67.4 + 6.86x$$

Alternatively, equation (14.10) may be solved by determinants. The formulas are:

$$(14.12) \quad b = \frac{\begin{vmatrix} N & \sum y \\ \sum x & \sum xy \end{vmatrix}}{\begin{vmatrix} \sum x & \sum xy \\ \sum x^2 & \sum x^2 \end{vmatrix}} \div D = \frac{N\sum xy - (\sum x)(\sum y)}{D}$$

$$(14.13) \quad a = \frac{\begin{vmatrix} \sum y & \sum x \\ \sum xy & \sum x^2 \end{vmatrix}}{\begin{vmatrix} \sum x & \sum xy \\ \sum x^2 & \sum x^2 \end{vmatrix}} \div D = \frac{(\sum x^2)(\sum y) - (\sum x)(\sum xy)}{D}$$

$$(14.14) \quad \text{where } D = \begin{vmatrix} N & \sum x \\ \sum x & \sum x^2 \end{vmatrix} = N\sum x^2 - (\sum x)^2$$

(The subscripts i have been dropped for convenience in printing.) It is understood, of course, that D is not zero. If it is, the equations for a and b are either incompatible (with no solutions) or equivalent to each other (with indeterminate solutions). This is not likely to happen in practice.

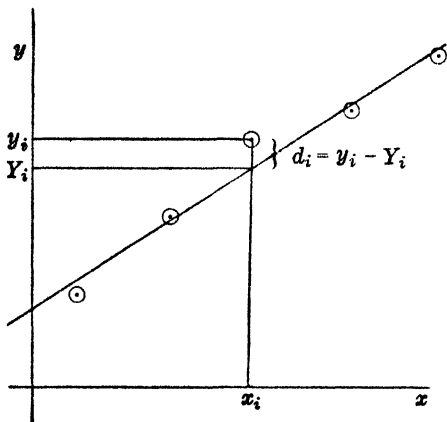


FIG. 52

14.9 The Method of Least Squares.

A second general method of fitting a mathematical curve to a set of data is known as the *method of least squares*. It depends on the principle that the "best" fit is obtained when the sum of squares of the differences d_i between the observed values y_i and the corresponding calculated values Y_i is as small as possible. (See Fig. 52.) That is,

$$(14.15) \quad \sum_{i=1}^N d_i^2 = \sum_{i=1}^N (y_i - Y_i)^2 = \text{minimum}$$

The differences d_i are called *residuals*. In some circumstances the d_i are weighted (see References 1 and 3 of the Introduction); but in fitting a straight line or polynomial to an ordinary time series the weights may all be taken as unity. For the straight line case, (14.15) becomes

$$(14.16) \quad \sum (y_i - a - bx_i)^2 = \text{minimum}$$

or, written out in full, with the subscripts dropped,

$$(14.17) \quad \sum y^2 + Na^2 + b^2 \sum x^2 - 2a \sum y - 2b \sum xy + 2ab \sum x = \text{minimum}$$

Now the left side of (14.17) can be regarded as a quadratic expression in b , of the form

$$(14.18) \quad Ab^2 + 2Bb + C$$

where $A = \sum x^2$, $B = a\sum x - \sum xy$, and $C = Na^2 - 2a\sum y + \sum y^2$.

Since A is certainly not zero (we cannot fit a curve to *one* point), equation (14.18) may be written as

$$\frac{1}{A} (A^2b^2 + 2ABb + AC) = \frac{1}{A} [(Ab + B)^2 + AC - B^2]$$

and this will have its minimum value for any choice of b when $Ab + B = 0$, that is, when

$$(14.19) \quad b\sum x^2 + a\sum x - \sum xy = 0$$

Again, (14.17) may be regarded as a quadratic in a , namely,

$$(14.20) \quad A'a^2 + 2B'a + C'$$

where $A' = N$, $B' = b\sum x - \sum y$, and $C' = b^2\sum x^2 - 2b\sum xy + \sum y^2$.

This has its minimum for any choice of a when $A'a + B' = 0$, that is, when

$$(14.21) \quad Na + b\sum x - \sum y = 0$$

These equations (14.19) and (14.21) may be obtained more simply by the student who knows a little calculus, by differentiating the lefthand side of (14.17) partially with respect to both a and b and putting the derivatives equal to zero. In any case, it is evident that these are the same equations as we obtained before (equations (14.10)) by the method of moments. It can be proved * that, for the fitting of any *polynomial* curve, the method of least squares and the method of moments lead to the same normal equations.

The first equation of (14.10) can be written

$$(14.22) \quad \sum d_i = 0$$

so that the sum of the residuals (taking account of the signs) is zero. This property, combined with (14.15), is analogous to the similar properties of the arithmetic mean, namely, that the sum of deviations from the mean is zero and the sum of squares of deviations from the mean is less than from any other value.

14.10 Fitting a Straight Line Through the Origin. If it is definitely known that the origin is a point on the curve, the line to be fitted is simply

$$Y = bx$$

and the least squares condition is

$$\sum (y - bx)^2 = \text{minimum}$$

or

$$(14.23) \quad b^2\sum x^2 - 2b\sum xy + \sum y^2 = \text{minimum}$$

* See Reference 2.

Applying the criterion above for a minimum, we have

$$b\sum x^2 - \sum xy = 0$$

or

$$b = \sum xy / \sum x^2$$

14.11 Simplification of Calculations for Equispaced Data. If, as usual in time series, the observations are taken at *equal intervals* of time, the calculation of the constants for fitted curves can be made much more easily. We first suppose that the number N of observations is odd ($= 2k + 1$). There is then a middle observation, the $(k + 1)$ th, which is the arithmetic mean of the x 's.

If the common time interval is c we can change to a new unit u , given by $u_i = (x_i - \bar{x})/c$, and in the new units the times of the observations are $-k, \dots, -3, -2, -1, 0, 1, 2, 3, \dots, k$. Clearly $\sum u_i = 0$, and $\sum u_i^2 = 2(1^2 + 2^2 + \dots + k^2) = k(k + 1)(2k + 1)/3 = l$, say. The normal equations (14.10), in terms of u , become

$$(14.24) \quad \begin{aligned} \sum y &= Na \\ \sum uy &= lb = b\sum u^2 \end{aligned}$$

whence

$$a = (\sum y)/N = \bar{y}, \quad b = (\sum uy)/l$$

The equation of the line is $Y = a + bu = a + b(x - \bar{x})/c$.

Example 3. For the following data, $c = 5$, $k = 2$, $l = 2 \cdot 3 \cdot 5/3 = 10$, $\bar{x} = 10$. (Instead of using the formula for l , we can work out the values of u^2 and add them.)

x	u	y	uy	u^2
0	-2	12	-24	4
5	-1	15	-15	1
10	0	17	0	0
15	1	22	22	1
20	2	24	48	4
		90	31	10

Then $a = 90/5 = 18$, $b = 31/10 = 3.1$. The equation is

$$\begin{aligned} Y &= 18 + 3.1u \\ &= 18 + \frac{3.1}{5}(x - 10) \\ &= 0.62x + 11.8 \end{aligned}$$

If the number of observations is even ($= 2k$), there is no middle value of x , but the arithmetic mean \bar{x} is midway between the two middle values. The values of u are fractional, but it is convenient to double them and compute $2uy$ and $4u^2$.

Example 4. The same data as in Example 3, but with an additional observation.

x	y	$2u$	$2uy$	$4u^2$
0	12	-5	-60	25
5	15	-3	-45	9
10	17	-1	-17	1
15	22	1	22	1
20	24	3	72	9
25	30	5	150	25
		<hr/>	<hr/>	<hr/>
		120	122	70

$\bar{x} = 12.5$, $k = 3$, $a = 120/6 = 20$, $b = 61/17.5 = 3.49$. The line is

$$\begin{aligned}
 Y &= 20 + 3.49u \\
 &= 20 + \frac{3.49}{5}(x - 12.5) \\
 &= 0.70x + 11.3
 \end{aligned}$$

For a long series a formula is useful for $\sum u^2 = \frac{1}{2}(1^2 + 3^2 + 5^2 + \cdots)$ for k terms. The formula is $m = \sum u^2 = N(N^2 - 1)/12$, where $N = 2k$. In Example 4, $k = 3$, $N = 6$, $m = 35/2 = 17.5$.

14.12 Exponential Trends. When the given y values form approximately a geometric progression while the corresponding x values form an arithmetic progression, the relationship between the variables is given by an exponential function, and the best fitting curve is said to describe an exponential trend. Data from the fields of biology, banking, and economics frequently exhibit such a trend. Thus the growth of bacteria is exponential. Money accumulating at compound interest follows the same kind of law of growth. And in business, sales or earnings may grow exponentially over a short period. Another familiar example is the increase in friction as a rope is coiled around a post. As the number of coils increases in arithmetic progression, the pull which the rope will stand without slipping increases in geometric progression.

The characteristic property of this law is that the rate of growth, that is, the rate of change of Y with respect to x , at any value of x , is proportional to the value of the function for that value of x . The function

$$(14.25) \quad Y = Ac^{Bx}, \quad A > 0$$

has this property.* The letter c is a fixed constant, usually either 10 or e , whereas A and B are statistics to be determined from the data. If Y decreases as x increases, B is negative. An interesting example of this case is the disappearance as time goes on of radioactive substances like radium.

To assume that the apparent law of growth will continue is usually unwarranted, so only short range predictions can be made with any considerable

* The student of calculus will understand that "rate of change" is used here in the sense of derivative. For (14.25), $dY/dx = kY$.

degree of reliability. When the exponential character of the observed phenomenon ceases a saturation point is said to be reached.

If we take logarithms (to base 10) of both sides of (14.25), we obtain

$$(14.26) \quad \log Y = \log A + (B \log c)x$$

If $c = 10$, $\log c = 1$; if $c = e$, $\log c = 0.4343$ approximately. In either case, (14.26) is of the form

$$(14.27) \quad Y' = a + bx$$

where

$$(14.28) \quad Y' = \log Y, \quad a = \log A, \quad b = B \log c$$

and is therefore the equation of a straight line in the coordinates Y' and x . A method of fitting an exponential trend line to a set of observed y 's is thus to fit a straight trend line to the *logarithms* of the y 's.

If we denote $\log y$ by y' , we have

$$(14.29) \quad \begin{cases} b = [N \sum xy' - (\sum x)(\sum y')]/D \\ a = [(\sum y')(\sum x^2) - (\sum x)(\sum xy')]/D \\ D = N \sum x^2 - (\sum x)^2 \end{cases}$$

Example 5. Find the exponential trend for the following data and draw the curve.

x	1	2	3	4	5
y	1.6	4.5	13.8	40.2	125.0

As before, the work can be shortened by using a new variable, $u = x - 3$. The necessary computations are

u	y	$y' = \log y$	uy'	u^2
-2	1.6	0.2041	-0.4082	4
-1	4.5	0.6532	-0.6532	1
0	13.8	1.1399	0	0
1	40.2	1.6042	1.6042	1
2	125.0	2.0969	4.1938	4
		<u>5.6983</u>	<u>4.7366</u>	<u>10</u>

Then

$$b = \sum uy' / \sum u^2 = 0.4737$$

$$a = \sum y' / N = 1.1397$$

Therefore

$$Y' = 1.1397 + 0.4737(x - 3)$$

(14.30)

$$= 0.4737x - 0.2814$$

If we want the form (14.25) we must put $A = \text{antilog } a = \text{antilog } (-0.2814) = 0.5231$; $B = 0.4737$, if $c = 10$, or $B = 0.4737/0.4343 = 1.091$, if $c = e$. The equation of the exponential trend is, therefore,

$$\begin{aligned} Y &= 0.5231(10)^{0.4737x} \\ (14.31) \quad &= 0.5231e^{1.091x} \end{aligned}$$

For the purposes of plotting the curve, predicting, or interpolating, it is usually best to calculate Y' from (14.30) and put $Y = \text{antilog } Y'$. Thus the plotted points in Fig. 53 are obtained from the following table:

x	1	2	3	4	5	6
Y'	0.1923	0.6660	1.1397	1.6134	2.0871	2.5608
Y	1.56	4.63	13.79	41.06	122.2	363.8

It should be noted that applying the least squares criterion to the fitting of the straight line (14.27) is not quite the same thing as applying it to the fitting of the exponential curve (14.25). Our method gives greater weight to the smaller values of y . This may be a reasonable thing to do, particularly with economic data, the estimated error in such data being often roughly proportional to the magnitude of the quantity measured. If, however, we feel that in an exponential time series the observations should all be weighted equally, we can achieve this result approximately by weighting the *logarithms* in proportion to y . The weighted least squares condition is *

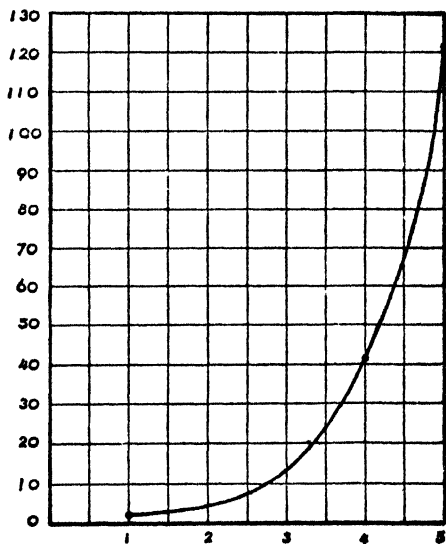


FIG. 53

$$(14.32) \quad \sum y d_i^2 = \text{minimum}$$

where

$$d_i = y_i' - Y_i' = y_i' - a - bx_i'$$

The corresponding normal equations are

$$(14.33) \quad a \sum y + b \sum xy = \sum yy', \quad a \sum xy + b \sum x^2 y = \sum xy y'$$

*See Part Two, page 318.

14.13 The Compound Interest Law. Equation (14.25) is sometimes called the compound interest law because it describes the way money would grow if interest were compounded regularly. If P dollars are invested at a nominal rate $100j\%$ compounded m times a year, the amount S dollars after x years is given by the formula

$$S = P \left(1 + \frac{j}{m} \right)^{mx}$$

If j is compounded continuously or, in other words, if m is taken indefinitely large (written $m \rightarrow \infty$), the amount S does not increase indefinitely but approaches a limiting value. We may write the expression for S in the form

$$S = P \left[\left(1 + \frac{j}{m} \right)^{m/j} \right]^{jx}$$

If we let $N = m/j$, we have

$$S = P \left[\left(1 + \frac{1}{N} \right)^N \right]^{jx}$$

It can be shown that, as $N \rightarrow \infty$, the quantity $\left(1 + \frac{1}{N} \right)^N$ approaches the limit called e . Thus we have

$$\lim_{N \rightarrow \infty} \left(1 + \frac{1}{N} \right)^N = e = 2.718 \dots$$

This limit is also the base of the Napierian, or natural, system of logarithms. As $m \rightarrow \infty$ so does $N \rightarrow \infty$. Therefore in the ideal case of continuous conversion of interest, we have the limiting form

$$\begin{aligned} S &= \lim_{m \rightarrow \infty} P \left[\left(1 + \frac{j}{m} \right)^{m/j} \right]^{jx} \\ &= \lim_{N \rightarrow \infty} P \left[\left(1 + \frac{1}{N} \right)^N \right]^{jx} \end{aligned}$$

that is

$$S = Pe^{jx}$$

which is of the form (14.25).

There are several other forms of the exponential function. For example, if we let $r = e^B$, (14.25) becomes

$$y = Ar^x$$

which is the general term of a geometric progression whose first term is A and common ratio is r .

If $c = e$, $Y = Ae^{Bx} = A10^{kx}$, where $k = B \log_{10} e = 0.4343 B$ approximately. The factor 0.4343 is called the *modulus* of logarithms to base 10. The reciprocal of the modulus, 2.3026, is often useful since for any number N

$$(14.34) \quad \log_e N = 2.3026 \log_{10} N$$

When the symbol \log is used without reference to base, we shall understand in the future that base 10 is meant. The symbol \ln is commonly used for a Napierian logarithm.

14.14 Semi-logarithmic Graph Paper. In the graphical representation of data that exhibit an exponential trend, it is often desirable to use semi-logarithmic paper. Such paper has a logarithmic scale in the vertical direction and a uniform scale in the horizontal direction (Fig. 54). A logarithmic scale is one in which the distance from $y = 1$ to $y = N$ equals $\log N$. A "cycle" of rulings spaced according to the logarithms of the integers from 1 to 10 is the unit of the vertical $\log y$ scale.

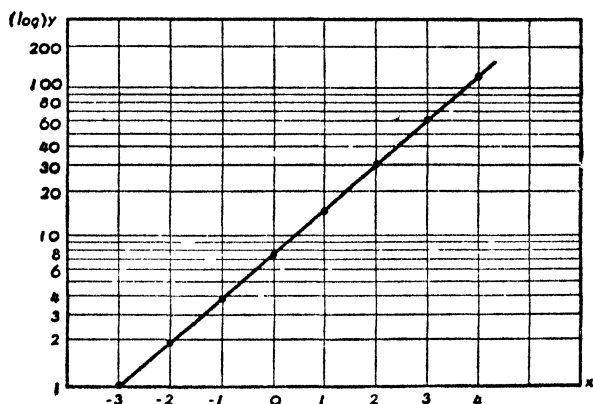


FIG. 54

"Semi-log" paper may be constructed or purchased having one or more cycles. The appropriate number of cycles is determined by the range of y values in the data to be plotted. If the bottom line of the first cycle is labeled 1 and taken as the origin of $\log y$ ($\log 1 = 0$), the beginning of the next cycle is read 10 ($\log 10 = 1$), the next one above that is read 100 ($\log 100 = 2$), etc. However, the beginning of the first cycle may be labeled with any number which is an integral power (positive or negative) of 10, as 0.01, 0.1, 10, 100, etc. Corresponding lines in successive cycles are labeled with numbers which are 10 times those in the preceding cycle. Since y has no real logarithm if $y \leq 0$, neither zero nor negative numbers are found on a logarithmic scale. Plotting a point whose semi-logarithmic coordinates are (x, y) is equivalent to plotting the point whose rectangular coordinates are $(x, \log y)$.

Example 6. Plot $y = 8(2^x)$ on semi-log paper.

Solution. Assigning values to x we form the following table,

x	-3	-2	-1	0	1	2	3	4
y	1	2	4	8	16	32	64	128

from which we obtain the straight-line semi-logarithmic graph shown in Figure 54.

Since (14.27) is linear in x and Y' , it is clear that (for $A > 0$) the graph of $Y = Ac^{Bx}$ on semi-logarithmic graph paper is bound to be straight. If, therefore, a time series is suspected to show an exponential trend, the simplest way to test this is to plot the points on semi-log paper and see whether they lie nearly on a straight line.

After drawing (by eye) what seems to be the best-fitting straight line, we can estimate roughly the constants of the exponential curve from the coordinates of two selected points on this line. Thus, if the line goes through the points (x_1, Y_1) and (x_2, Y_2) , the slope is given by

$$b = \frac{Y_2' - Y_1'}{x_2 - x_1} = \frac{\log(Y_2/Y_1)}{x_2 - x_1} = B \log c$$

If we choose the points x_1 and x_2 so that the interval corresponds to one cycle,

$Y_2/Y_1 = 10$, and therefore $B \log c = \frac{1}{x_2 - x_1}$, or, when $c = e$,

$$B = 2.3026/(x_2 - x_1)$$

The value of A can be estimated from the point where the straight line cuts $x = 0$. If this point is not on the graph, we can choose any convenient x , note the corresponding Y , and calculate $A = Ye^{-Bx}$, from a table of $e^{-\lambda}$. (See Table 37, §10.8.)

14.15 Ratio Charts. Graphs on semi-log paper are often called ratio charts. Their usefulness depends upon the property of logarithms that

$$\log \frac{M}{N} = \log M - \log N$$

It follows that the distance between any two ordinates of the chart measures the ratio between the values represented by these ordinates. Thus if

$$\frac{y_1}{y_3} = \frac{y_2}{y_4}$$

then

$$\log y_1 - \log y_2 = \log y_3 - \log y_4$$

or

$$Y_1 - Y_2 = Y_3 - Y_4$$

that is, equal ratios are represented by equal vertical distances. Likewise, if

$$\frac{y_1}{y_2} > \frac{y_3}{y_4}$$

then

$$Y_1 - Y_2 > Y_3 - Y_4$$

and the larger ratio is represented graphically by the larger distance. These differences of elevation are independent of any base line. The same percentage increase in y is represented by the same addition to the height of Y in all parts of the chart. Hence, it is easier to depict and discover percentage changes on ratio charts than on ordinary charts.

The analysis of time series in economic statistics is often facilitated by forming "link relatives" which are ratios of each ordinate (after the first) to the preceding ordinate. Thus, if y_1, y_2, \dots, y_n are the given values, the link relatives are

$$R_1 = \frac{y_2}{y_1}, \quad R_2 = \frac{y_3}{y_2}, \dots, \quad R_{n-1} = \frac{y_n}{y_{n-1}}$$

For any link relative, $100(R - 1)$ denotes the percentage change in y from one month (say) to the next. If the y 's are plotted on ratio paper they will lie on a straight line when the R 's are equal, on a curve bending upward when the R 's are increasing, and on a curve bending downward when the R 's are decreasing. It follows that if two curves are parallel on ratio paper their percentage rate of increase (or decrease) is the same.

TABLE 55. DEATH RATES PER 100,000 (U. S. A.) FOR TUBERCULOSIS AND TYPHOID FEVER, 1900-1920

Year	Tuberculosis	Typhoid	Year	Tuberculosis	Typhoid
1900	195.2	31.3	1911	159.0	15.3
1901	189.8	27.5	1912	149.8	13.2
1902	174.1	26.3	1913	148.7	12.6
1903	177.1	24.6	1914	148.6	10.8
1904	188.5	23.9	1915	146.7	9.2
1905	180.9	22.4	1916	143.8	8.8
1906	177.8	22.0	1917	147.1	8.1
1907	175.6	20.5	1918	151.0	7.0
1908	169.4	19.6	1919	124.9	4.8
1909	163.3	17.2	1920	112.0	5.0
1910	164.7	18.0			

An example of the different impressions that may be given by plotting the same observations on ordinary and on semi-log paper is furnished by the data in Table 55, relating to the death rates in certain states of the U.S.A. for

tuberculosis (all forms) and for typhoid fever. In Fig. 55 these data are plotted on ordinary graph paper and in Fig. 56 on semi-log graph paper. The decline in absolute value is greater for T.B. than for typhoid, as shown by the steeper slope in Fig. 55, but the *relative* decline (the percentage decrease)

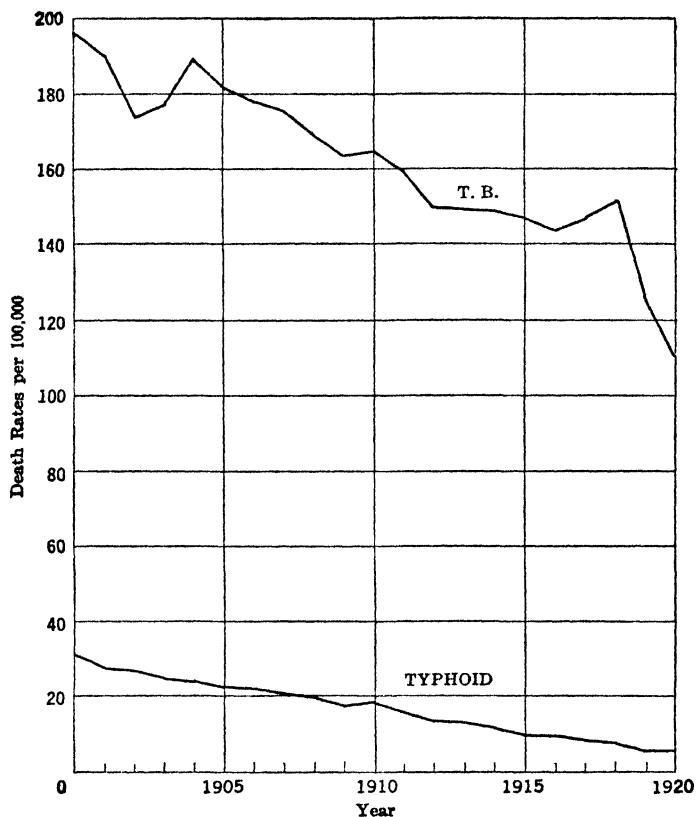


FIG. 55. DEATH RATES FROM TUBERCULOSIS AND TYPHOID FEVER

is greater for typhoid than for T.B. (about 80% as against 40%), as shown by the steeper slope in Fig. 56. The two graphs bring out different aspects of the same statistical data.

14.16 Logarithmic Graph Paper. Coordinate paper on which the rulings in *both* directions are at distances from the origin proportional to the logarithms of the numbers represented is called logarithmic paper or log-log paper. It is not used very much for time series, but is mentioned here because of the similarity to semi-log paper, which is used a great deal.

Its main purpose is to represent by a straight line *power functions* of the form

$$(14.35) \quad Y = Ax^b, \quad A > 0$$

which can be written

$$\log Y = \log A + b \log x$$

or

(14.36)

$$Y' = a + bx'$$

where

$$x' = \log x, \quad Y' = \log Y, \quad a = \log A.$$

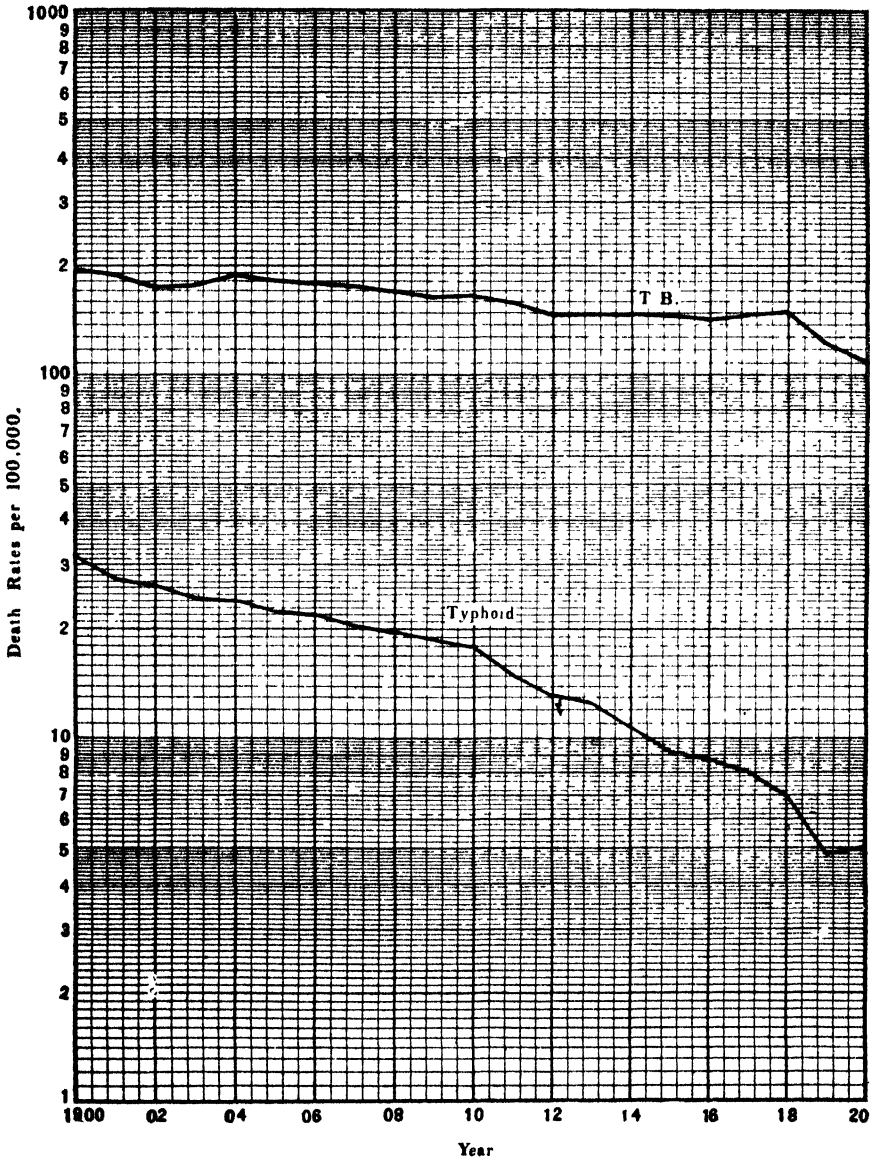


FIG. 56. DEATH RATES FROM TUBERCULOSIS AND TYPHOID FEVER

Since on log-log paper the abscissas and ordinates are proportional to x' and Y' , the graph of (14.36) on such paper is a straight line. The constants a and b can be approximately evaluated from the slope of this line and its intercept on the Y axis. More exact values can be obtained by fitting the straight line (14.36) by least squares to the *logarithms* of the observed x and y values.

Log-log paper is useful for graphing distributions which cover a very great range in both variables. A table of cumulative frequencies for persons in the U.S. reporting incomes in excess of a specified amount is of this nature. The x -variable (income) may range from \$2000 up to a million or more, and the y -variable (cumulative frequency) from a few tens up to a few millions. The data would be difficult to plot except on log-log paper covering several cycles.

This kind of graph paper is also frequently useful in engineering, where empirical relationships of the power function type often occur. The relation between pressure and volume of a gas in adiabatic expansion, $p = kv^{-\gamma}$, is of this type, and so is the formula for the flow of water over a rectangular weir of breadth B and height H , $Q = 3.33 BH^{3/2}$. Both these relationships give straight lines on log-log paper.

14.17 Other Types of Trend. The fitting of a parabolic trend line by least squares will be described in Chapter 16. Some other types of trend line occasionally required are:

- (a) the modified exponential curve,
- (b) the Gompertz curve,
- (c) the Makeham curve,
- (d) the logistic.

Modified Exponential Curve. This curve has the equation

$$(14.37) \quad Y = A + Be^{px} = A + Bq^x, \quad q = e^p$$

and the graph on semi-log paper is concave upward if A is positive and concave downward if A is negative. (B is assumed positive. The curvatures are reversed if B is negative.) A graphical method of fitting, due to Cowden (Reference 3), consists in plotting the data on ordinary or semi-log graph paper, drawing a tentative trend line and selecting three equidistant ordinates, well apart, say at $x - c$, x , and $x + c$. If the corresponding ordinates are Y_0 , Y_1 and Y_2 , we have from (14.37)

$$(14.38) \quad \begin{cases} Y_0 = A + Be^{p(x-c)} \\ Y_1 = A + Be^{px} \\ Y_2 = A + Be^{p(x+c)} \end{cases}$$

Therefore

$$(14.39) \quad (Y_2 - Y_1)/(Y_1 - Y_0) = (e^{pc} - 1)/(1 - e^{-pc}) = e^{pc}$$

Also

$$\begin{aligned} Y_1 - Y_0 &= Be^{px}(1 - e^{-pc}) \\ &= Be^{p(x-c)}(e^{pc} - 1) \end{aligned}$$

$$\text{From (14.39),} \quad e^{pc} - 1 = \frac{Y_2 - 2Y_1 + Y_0}{Y_1 - Y_0}$$

$$\text{so that} \quad Be^{p(x-c)} = (Y_1 - Y_0)^2 / (Y_2 - 2Y_1 + Y_0)$$

Then, from the first equation of (14.38),

$$\begin{aligned} (14.40) \quad A &= Y_0 - (Y_1 - Y_0)^2 / (Y_2 - 2Y_1 + Y_0) \\ &= (Y_0Y_2 - Y_1^2) / (Y_2 - 2Y_1 + Y_0) \end{aligned}$$

Having estimated A in this way we plot values of $y_t - A$ (the y_t are the observed values) on semi-log paper. If necessary we adjust A a little until a straight line fits reasonably well. Then the ordinate of this line at $x = 0$ gives the value of B and the ratio of the ordinates at x and $x - c$ gives $e^{pc} = q^c$, from which q can be found. If the values of $y_t - A$ are negative, the sign of the scale values on the semi-log paper must be changed.

Gompertz Curve. Type (b) is used in actuarial work and has had some application as a growth curve in business and population forecasting. Its equation* is

$$(14.41) \quad Y = ab^{q^x}$$

or, in logarithmic form,

$$(14.42) \quad \log Y = \log a + q^x \log b = A + Bq^x$$

where $A = \log a$, $B = \log b$. This is the same equation as (14.37) with $\log Y$ instead of Y . If $q < 1$, we see from (14.42) that $Y \rightarrow a$ as $x \rightarrow \infty$. The line $Y = a$ is an asymptote to the curve, and a is sometimes called the *ceiling* of the curve. (Fig. 57.)

The curve may be fitted by a modification of Cowden's method described before, plotting $\log y_t$ instead of y_t .

Makeham Curve. Type (c) is also used in actuarial work. The equation is

$$(14.43) \quad Y = ks^xb^{q^x}$$

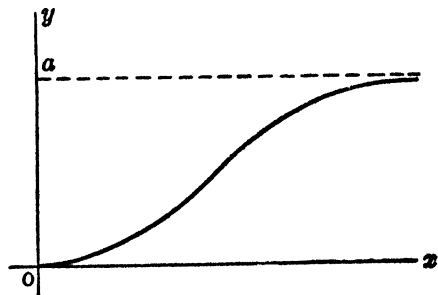


FIG. 57

*For a derivation see References 4 or 5.

or

$$(14.44) \quad \begin{aligned} \log Y &= \log k + x \log s + q^x \log b \\ &= A + Bq^x + Cx, \quad C = \log s \end{aligned}$$

It is a combination of a straight line with a Gompertz curve. For a discussion of its use in the field of insurance, see Reference 5.

Logistic. This curve has been widely used in the study of growth. Its equation is

$$(14.45) \quad Y = a/(1 + bq^x)$$

or

$$(14.46) \quad \begin{aligned} \frac{1}{Y} &= \frac{1}{a} + \frac{b}{a} q^x \\ &= A + Bq^x \end{aligned}$$

which is the same as (14.37) with $1/Y$ instead of Y . The same method of fitting will apply, with the reciprocals $1/y_i$ plotted instead of the y_i . For convenience of computing trend values when the constants have been determined, (14.46) may be written

$$(14.47) \quad \log \left(\frac{1}{Y} - A \right) = \log B + x \log q$$

A logistic curve is a fairly good fit to the curve of population of the U.S.A. shown in Fig. 49, §14.1. Recent census figures, however, have not followed the logistic trend as closely as was hoped by Raymond Pearl and others who first advocated the use of this curve for population studies.

14.18 The Analysis of Business Time Series. Long series of monthly data, such as those on the mineral production of the United States over twenty or thirty years, or the sales of a large business concern over a similar time, may be analyzed into (a) a long-term trend, T ; (b) oscillations or cycles, C ; (c) seasonal fluctuations, S ; (d) random irregular variations, I ; each of these may be regarded as proportional to the preceding ones. That is to say, C fluctuates around the trend as a base, S fluctuates around a curve combining trend and cycles, while I fluctuates around the combined curve of trend, cycles, and seasonal variation, and the combined effect $T + C + S + I$ may be written

$$T \left(\frac{T + C}{T} \right) \left(\frac{T + C + S}{T + C} \right) \left(\frac{T + C + S + I}{T + C + S} \right)$$

The second factor may be regarded as the cyclical effect, expressed as a fraction of T and fluctuating around the value 1. We denote it by C_1 , and simi-

larly the other factors may be denoted by S_1 and I_1 . Then the time series is expressed by $y = TC_1 S_1 I_1$, and consequently

$$\log y = \log T + \log C_1 + \log S_1 + \log I_1$$

This means that if the various components are plotted separately on *semi-log paper*, the ordinate of y on the same paper will be the *sum* of the ordinates of the separate components.

The analysis of the time series into these components involves several steps:

1. Estimating seasonal movements.
2. Adjusting the data by dividing by the seasonal index S_1 .
3. Computing the trend either by moving averages or by a mathematical equation.
4. Adjusting for trend by dividing by T .
5. Smoothing out the irregularities by a short moving average. This leaves only the cyclical movements.

14.19 Deflating and Deseasonalizing Data. Before attempting to analyze a series expressing business activity it is often advisable to see that the figures are as comparable with each other as possible. For example, if one is interested in the variation in volume of sales of a commodity, the dollar value of the sales might not be a true indication of this variation because of the change in price through the period studied. Hence the figures in a value series are often divided each by an appropriate price index number to obtain a comparable quantity series, representing volume (of sales, production, etc.). This process is known as *deflating* the series.

Another difficulty is caused by the irregularities of the calendar, in particular by the different lengths of the months. Monthly sales figures may be rendered more comparable by multiplying them by a factor which is the ratio of the average number of days in a month to the actual number in a particular month. Production figures for an industry may be adjusted by a factor based on the number of working days in the month. We suppose adjustments such as these carried out before the analysis of the time series begins. The first step is to smooth out the seasonal movements by a 12-month moving average (usually followed by a 2-month moving average, so as to center the final average on a calendar month). The resulting curve is an estimate of the trend and the cyclical movements combined (TC). The original data are divided by these TC estimates, giving estimates of SI (seasonal and irregular movements), corrected for the effects of trend and cycles, and expressed as percentages of the centered 12-month moving average.

An illustration of this process is furnished by the fictitious data of Table 56, relating to sales of the XYZ Products Company. The monthly figures

are in column (2), the centered 12-month moving average in column (5), and the seasonal and irregular movements, as a percentage of the moving average, in column (6). Columns (2) and (5) are graphed in Fig. 58 and column (6) in Fig. 59.



FIG. 58. SALES AND MOVING AVERAGE

A seasonal index is then formed by averaging all the January values, all the February values, and so on, in column (6), and adjusting the results by a suitable factor so that the twelve averages themselves average exactly 100 per cent. For brevity, Table 56 has been confined to 3 years, and there are only two values for each month in column (6), but in practice there would probably be a series covering 15 or 20 years to work on. However, to illustrate the method we will compute the seasonal index from the limited data given.

TABLE 56. MONTHLY SALES OF XYZ PRODUCTS CO.

(1) Date		(2) Sales (thousands of dollars)	(3) 12-mo. moving total	(4) 2-mo. moving total	(5) Centered 12-mo. mov- ing average ((4) + 24)	(6) Estimated SI (100(2) + (5))
1949	J	3639				
	F	3591				
	M	3326				
	A	3469				
	My	3321				
	J	3320				
			41 424			
	Jy	3205		83 122	3463	92.5
			41 698			
	A	3205		83 661	3486	91.9
			41 963			
	S	3255		84 314	3513	92.7
1950			42 351			
	O	3550		85 053	3544	100.2
			42 702			
	N	3771		85 730	3572	105.6
			43 028			
	D	3772		86 234	3593	105.0
			43 206			
	J	3913		86 683	3612	108.3
			43 477			
	F	3856		87 103	3629	106.3
			43 626			
	M	3714		87 591	3650	101.8
			43 965			
	A	3820		88 210	3675	103.9
			44 245			
	My	3647		88 902	3704	98.5
			44 657			
	J	3498		90 024	3751	93.3
			45 367			
	Jy	3476		91 214	3801	91.4
			45 847			
	A	3354		92 368	3849	87.1
			46 521			
	S	3594		93 615	3901	92.1
			47 094			
	O	3830		94 773	3949	97.0
			47 679			
	N	4183		95 735	3989	104.9
			48 056			
	D	4482		96 606	4025	111.4
			48 550			

TABLE 56. MONTHLY SALES OF XYZ PRODUCTS Co. — Continued

(1) Date	(2) Sales (thousands of dollars)	(3) 12-mo. moving total	(4) 2-mo. moving total	(5) Centered 12-mo. mov- ing average ((4) + 24)	(6) Estimated SI (100(2) ÷ (5))
1951 J	4393		97 419	4059	108.2
F	4530	48 869	97 876	4078	111.1
M	4287	49 007	97 991	4083	105.0
A	4405	48 984	98 061	4086	107.8
My	4024	49 077	97 955	4081	98.6
J	3992	48 878	97 154	4048	98.6
Jy	3795	48 276			
A	3492				
S	3571				
O	3923				
N	3984				
D	3880				

	J	F	M	A	My	J	Jy	A	S	O	N	D
	108.3	106.3	101.8	103.9	98.5	93.3	92.5	91.9	92.7	100.2	105.6	105.0
	108.2	111.1	105.0	107.8	98.6	98.6	91.4	87.1	92.1	97.0	104.9	111.4
Av.	108.2	108.7	103.4	105.8	98.6	96.0	92.0	89.5	92.4	98.6	105.2	108.2

If one or two out of a dozen or more monthly values differ markedly from the rest, they may be excluded in forming the average, as the purpose of this average is to get a typical representative number for the month.

The total of the averages in our example is 1206.6 instead of 1200, so each of them is multiplied by the factor $1200/1206.6 = 0.9945$. The final seasonal index is

J	F	M	A	My	J	Jy	A	S	O	N	D
107.6	108.1	102.8	105.2	98.1	95.5	91.5	89.0	91.9	98.1	104.6	107.6

The original data are *deseasonalized* by dividing each month's figure by the seasonal index for that month.

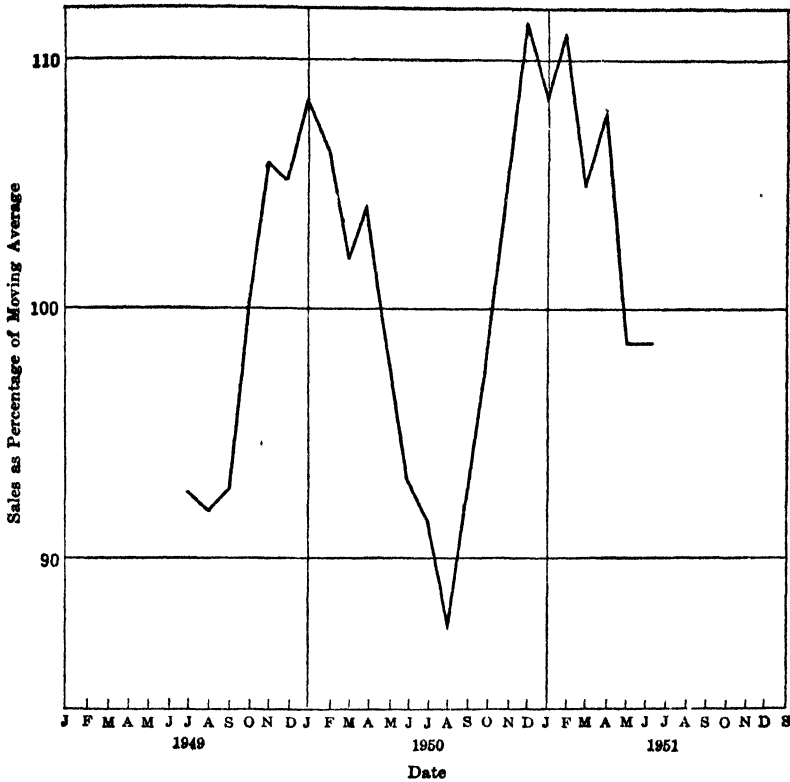


FIG. 59. SEASONAL AND IRREGULAR COMPONENTS

14.20 Elimination of Trend and Irregularities. The type of trend curve to be fitted depends on the appearance of the data as plotted, and no rules can be given. If possible, one that can be expressed by a mathematical equation should be used, and, unless there are theoretical or logical reasons for preferring a certain equation, the simpler the better. By way of illustration we have fitted two straight lines to the deseasonalized data of Table 56. The computations are shown in Table 56a. When the data are divided by the trend values, the result is the cyclical and irregular component only (*CI*). The irregularities may be smoothed out by taking a short moving average, such as a binomially weighted average of 3. The cyclical components then remain, and their interpretation (with real data) constitute the task of the economist. In Fig. 60 the fit of the straight trend lines to the deseasonalized data is indicated. The cyclical and irregular components may be plotted separately, but their general appearance can be estimated from the figure.

TABLE 56a. FIT OF TWO STRAIGHT TREND LINES TO DESEASONALIZED DATA

Date	Deseasonalized Sales (<i>y</i>) (thousands of dollars)	<i>u</i>		<i>uy</i>		$Y = a + bu$	
		(1)	(2)	(1)	(2)	(1)	(2)
1949 J	3382	-13		-48 966		3250	
F	3322	-12		-39 864		3280	
M	3235	-11		-35 585		3315	
A	3298	-10		-32 980		3348	
My	3385	-9		-30 465		3380	
J	3476	-8		-27 808		3418	
Jy	3503	-7		-24 521		3446	
A	3601	-6		-21 606		3478	
S	3542	-5		-17 710		3511	
O	3619	-4		-14 476		3544	
N	3605	-3		-10 815		3676	
D	3506	-2		- 7 012		3609	
1950 J	3637	-1		- 3 637		3642	
F	3567	0		0		3674	
M	3613	1		3 613		3707	
A	3631	2		7 262		3740	
My	3718	3		11 154		3772	
J	3663	4		14 652		3805	
Jy	3799	5		18 995		3838	
A	3769	6		22 614		3870	
S	3911	7		27 377		3903	
O	3904	8		31 232		3936	
N	3999	9		35 991		3968	
D	4165	10		41 650		4001	
1951 J	4083	11		44 913		4034	
F	4191	12	-5	50 292	-20 955	4066	4274
M	4170	13	-4	54 210	-16 680	4099	4223
A	4187		-3		-12 561		4172
My	4102		-2		- 8 204		4121
J	4180		-1		- 4 180		4069
Jy	4148		0		0		4018
A	3924		1		3 924		3967
S	3886		2		7 772		3916
O	3999		3		11 997		3865
N	3809		4		15 236		3814
D	3606		5		18 030		3763

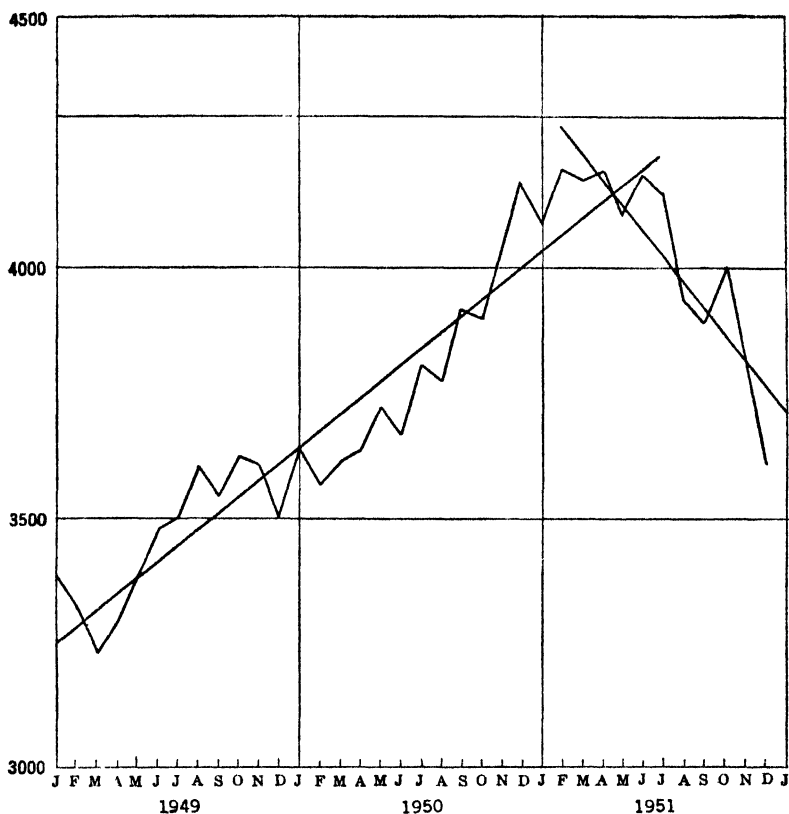


FIG. 60. DESEASONALIZED SALES DATA AND TREND LINES

Exercises

1. (*Wilson and Tracy*) The premium (\$ y) on a \$1000 life insurance policy for various ages (x years) is given in the following table. Draw a graph exhibiting y as a function of x . Estimate from the graph the premium at age 32 and at age 43; also the age at which the premium is \$52.

x	20	25	30	35	40	45	50	55	60
y	18.78	21.02	23.86	27.54	32.36	38.83	47.68	59.88	76.94

2. Find the equation of each of the straight lines through two points given as follows: (a) (2, 6), (4, 5); (b) (0, 3), (1, 6).

3. Find the equation of a line through the point (2, 3) and parallel to the line $4x + 5y = 7$.

4. Find the value of x for which $f(x) = 2x^2 - 8x + 9$ has a minimum value. What is this minimum?

Hint. Use the method of §14.9.

5. Fit a straight line to the following data:

x	6	7	8	9	10	11	12	13
y	7	7	6	4	4	3	2	1

Ans. $Y = -0.90x + 12.8$.

6. Calculate the values of Y for each value of x in Exercise 5, obtain the values of the deviations d_i , and check that $\sum d_i = 0$.

7. Calculate $\sum d_i^2$ in Ex. 6. Another line for which $\sum d_i = 0$ is $Y = -x + 13.75$. Check that $\sum d_i^2$ for this line is greater than for $Y = -0.90x + 12.8$.

8. The uniform horizontal scale on a sheet of semi-log paper ranges from 0 to 10. The vertical logarithmic scale ranges from 100 to 1000. A straight line is drawn on the paper from the upper end point of the vertical scale to the mid-point of the horizontal scale. Find the exponential function represented by the line. What is the equation of the line in (Y', x) coordinates, where $Y' = \log_{10} Y$? *Ans.* $Y = 1000 e^{-0.46x}$; $Y' = 3 - 0.2x$.

9. A straight line is drawn on logarithmic graph paper through the points (4, 16) and (6, 54). Find the function represented by this line. *Ans.* $y = x^3/4$.

10. Draw the graph of $y = 10 e^{-2x}$ on semi-log paper.

11. In the following table y represents the fire losses in the United States for the years mentioned (in millions of dollars). Find the best fitting straight line (in the least squares sense) for the data.

x	1915	1917	1919	1921	1923	1925
y	172	290	321	495	535	570

12. Add the pair of values $x = 6$, $y = 300$ to the data of Example 5, §14.12, and find the equation of the best fitting exponential curve. *Ans.* $Y' = 0.4617x - 0.2534$,
 $Y = 0.56e^{1.06x}$.

13. Sketch the curves $y = 10 e^{-x}$ and $y = 10 e^{-x^2/2}$, for $0 \leq x \leq 3$.

14. A straight line is drawn on semi-log paper through the points (2, 1) and (4, 100). What function has this line for its graph? *Hint.* Put $Y = Ar^x$. *Ans.* $100Y = 10^x$.

15. Data from a certain experiment involving voltage (v) as a function of time (t) are plotted on logarithmic coordinate paper, and are found to exhibit a linear trend there. A line is drawn, with a transparent ruler, which seems to fit the plotted data best. Two points on this line are (6, 18) and (8, 32). Determine an equation expressing v in terms of t whose logarithmic graph is the line.

16. Draw the graph of $y = 25x^n$ on logarithmic coordinate paper, (a) when $n = 2$, (b) when $n = -2$. Mark scales clearly.

17. Fit a logistic curve to the population figures of the United States, 1790–1950, using the method described in §14.17. The approximate figures, in millions, are:

x	1790	1800	1810	1820	1830	1840	1850	1860	1870	1880
y	3.93	5.31	7.24	9.64	12.87	17.07	23.19	31.44	39.82	50.16
x	1890	1900	1910	1920	1930	1940	1950			
y	62.95	76.00	91.97	105.71	122.78	131.67	150.70			

References

1. A. C. Aitken, *Statistical Mathematics*. (Oliver and Boyd, 5th ed., 1947.)
2. Dunham Jackson, "The Method of Moments," *Amer. Math. Monthly*, **30**, 1923, pp. 307-311.
3. D. J. Cowden, "Simplified Methods of Fitting Certain Types of Growth Curves," *J. Amer. Stat. Assoc.*, **42**, 1947, pp. 585-590.
4. C. H. Forsyth, *Mathematical Theory of Life Insurance* (John Wiley & Sons, Inc., 1924), pp. 67-68.
5. H. L. Rietz, "On Certain Properties of Makeham's Law of Mortality with Applications," *Amer. Math. Monthly*, **28**, 1921, pp. 158-165.
6. For fuller details concerning the treatment of business statistics, see Croxton and Cowden, *Practical Business Statistics*, 2nd edition (Prentice-Hall, Inc., 1948).

CHAPTER XV

LINEAR REGRESSION AND CORRELATION

15.1 Bivariate Data. Until now we have been concerned with problems of variation in a single variable quantity, or (in Chapter XIV) with a variable quantity depending on the time. We shall now consider the simultaneous variation of *two* variable quantities. The methods of expressing the relationship between two variables are due mainly to the English biometricians Sir Francis Galton (1822–1911) and Karl Pearson (1857–1936).

Data presenting two sets of related measurements or observations may arise in many fields of activity yielding N pairs of corresponding observations $(x_i, y_i), i = 1, 2, 3, \dots, N$. Thus x may represent July rainfall and y the average yield of corn in a certain section; x may be an index of commodity prices and y an index of employment over the same period; we may be interested in a group of school children in which x is their height and y their weight, or x may refer to their reading ability and y to their spelling ability; we may be studying the chance distributions which are obtained in throwing two dice where x is the number obtained in throws of a single die and y is the number obtained in throws of the two dice together.

Example 1. In the following set of selected heights (inches), x = stature of father, y = stature of son.

x	69	70	69	68	70	73	69	67	69	64
y	68	69	72	67	70	71	72	66	71	65

Example 2. (*Snedecor*) The following data on twelve trees are adapted from the results of an experiment to test the phenomenon that the injury by codling moth larvae seems to be greatest on apple trees bearing a small crop. Here x = hundreds of fruit on a tree, y = percentage of fruits wormy.

x	15	15	12	26	18	12	8	38	26	19	29	22
y	52	46	38	37	37	37	34	25	22	22	20	14

When the given pairs of values (x_i, y_i) are plotted on ordinary graph paper, we obtain a "dot diagram" or "scatter diagram." Fig. 61 shows the scatter diagram for the data of Example 2. There are two main problems

involved in the relationship between x and y . The first is to find the most suitable form of equation for use in predicting the average y for a given x or in predicting the average x for a given y , and to estimate the error in such predictions. This is the problem of *regression*. The second is to find a measure of the degree of association, or *correlation* as it is called, between the values of x and those of y . The two problems are closely related.

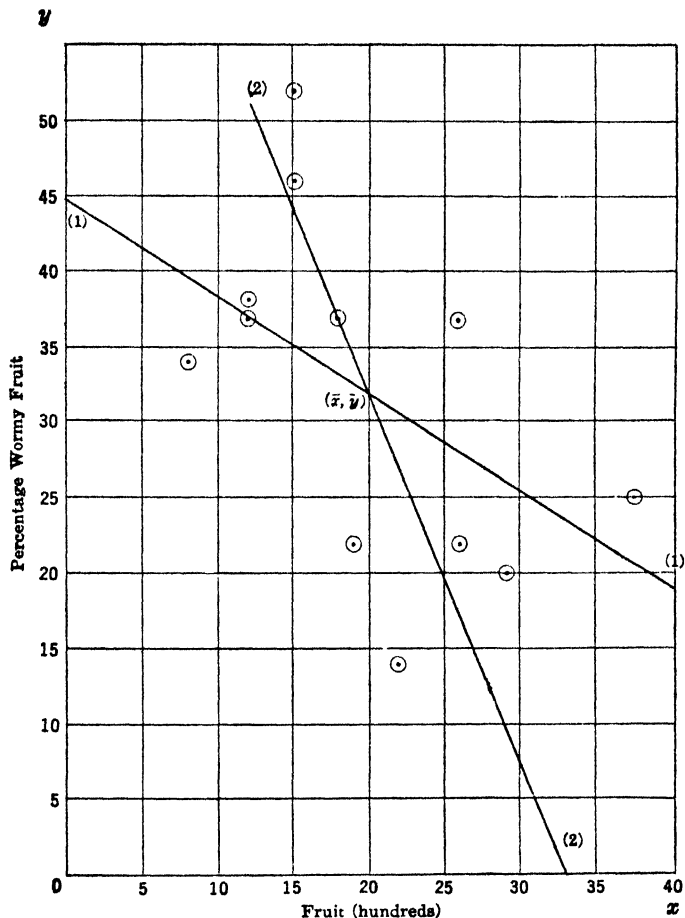


FIG. 61

The field of relationship may be thought of as bounded on the one extreme by perfect functional dependence and on the other extreme by complete independence in the probability sense. For example, the pairs of values which satisfy the equation $y = 2x - 5$ do not present a statistical problem. In this case the relationship is defined by a mathematical function $y = f(x)$.

Similarly, at the other extreme we would not be concerned with pairs of values which are completely independent in the probability sense, as, for example, the grades of students in statistics and the heights of their fathers. Two variables are said to be statistically related when they lie between these two extremes of relationship.

15.2 Regression. If we fit a straight line by least squares to the dots of the scatter diagram in such a way as to minimize the sum of the squares of the distances parallel to the y axis from the dots to the line, we obtain the *regression line of y on x* . As we saw in §14.9 this line has the equation

$$(15.1) \quad Y = a + bx$$

where a and b are given by

$$(15.2) \quad \begin{aligned} Na + \sum xb &= \sum y \\ \sum xa + \sum x^2b &= \sum xy \end{aligned}$$

We cannot simplify these equations as we did for the trend line fitted to a time series, because the x values are not usually equally spaced. The same general solution holds, however, namely:

$$(15.3) \quad b = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}$$

$$(15.4) \quad a = (\sum y - b \sum x)/N$$

The quantity b , which is the slope of the regression line, is usually called the *regression coefficient*.

Now, if s_x^2 is the variance of the N values x_i , we have

$$(15.5) \quad Ns_x^2 = \sum (x - \bar{x})^2, \quad \bar{x} = (\sum x)/N$$

and in a similar way we define a quantity s_{xy} , called the *covariance* of the N pairs of values x_i, y_i , by the relationship

$$(15.6) \quad Ns_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

Unlike the variance, this quantity may be positive or negative. Just as we may write

$$(15.7) \quad Ns_x^2 = \sum x^2 - (\sum x)^2/N$$

so we have

$$(15.8) \quad Ns_{xy} = \sum xy - (\sum x)(\sum y)/N$$

which is easily proved to be equivalent to (15.6). With this notation, (15.3) and (15.4) become

$$(15.9) \quad b = s_{xy}/s_x^2, \quad a = \bar{y} - b\bar{x}$$

so that the equation of the regression line is

$$(15.10) \quad Y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x})$$

The line, therefore, passes through the point whose coordinates are (\bar{x}, \bar{y}) .

The term *regression* was used first by Galton in studying inheritance of stature. He found that offspring of abnormally tall or short parents tend to "step back" or "regress" to the ordinary population height. However, as now used, regression line has no reference to biometry, but is merely a convenient term.

For the actual calculation of the constants a and b , with a set of discrete data, it is generally best to use (15.3) and (15.4). The computation for the data of Example 2 is set out in Table 57.

TABLE 57. CALCULATION OF VARIANCES AND COVARIANCE FOR DATA OF EXAMPLE 2

x	y	x^2	y^2	xy
15	52	225	2704	780
15	46	225	2116	690
12	38	144	1444	456
26	37	676	1369	962
18	37	324	1369	666
12	37	144	1369	444
8	34	64	1156	272
38	25	1444	625	950
26	22	676	484	572
19	22	361	484	418
29	20	841	400	580
22	14	484	196	308
<hr/>				
240	384	5708	13716	7098

We have

$$b = \frac{12(7098) - (240)(384)}{12(5708) - (240)^2} = \frac{-6984}{10896}$$

$$= -0.641$$

$$a = [384 - 240(-0.6410)]/12 = 44.82$$

so that the line is

$$(15.11) \quad Y = 44.82 - 0.641x$$

This line is marked (1) in Fig. 61. It is the appropriate line to use if we wish to estimate y for a given value of x . The x values are considered to be without appreciable error and the y values to vary in a random manner around the regression line.

A second regression line, that of x on y , may be fitted so as to minimize the sum of squares of the *horizontal* distances (parallel to the x axis) from the points to the line. This line has the equation

$$(15.12) \quad X - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y})$$

and so, like the other, passes through the point (\bar{x}, \bar{y}) . Its equation may be written

$$(15.13) \quad X = a' + b'y$$

where

$$(15.14) \quad b' = (N \sum xy - \sum x \sum y) / (N \sum y^2 - (\sum y)^2) = s_{xy} / s_y^2$$

$$(15.15) \quad a' = (\sum x - b' \sum y) / N$$

and this form is convenient for calculation. From the data of Table 57 we find

$$b' = [12(7098) - 240(384)] / [12(13716) - (384)^2]$$

$$= -6984 / 17136 = -0.4076$$

$$a' = [240 - 384(-0.4076)] / 12 = 33.04$$

$$X = 33.04 - 0.4076y$$

This line, marked (2) in Fig. 61, is the appropriate regression line to use in estimating x for a given value of y , that is, when the y values may be considered free from error and the x values as scattering about the regression line. Since, however, we are free to choose which of our variables shall be called x and which y , and since there is generally one variable which it is reasonable to regard as dependent on the other, it is not necessary to use both regression lines.

15.3 Coefficient of Correlation. There is said to be positive correlation between x and y if, for an assigned x greater than \bar{x} , the corresponding y values tend to be greater than \bar{y} , and if, for x less than \bar{x} , the corresponding y values tend to be less than \bar{y} . The correlation is negative if, for $x > \bar{x}$, y tends to be less than \bar{y} and if, for $x < \bar{x}$, y tends to be greater than \bar{y} . This definition depends on the assumption that the observed pairs of values (x_i, y_i) form a sample from an indefinitely large population in which y is a random variable having a probability distribution depending on x . If the sample is large, the observed y_i will give a good idea of this probability distribution.

When the variables are correlated there is a tendency for the dots in the scatter diagram to fall into a sort of band having a fairly definite trend. We are assuming that this trend is linear, and a theory built upon this assumption is known as simple or *linear* correlation.

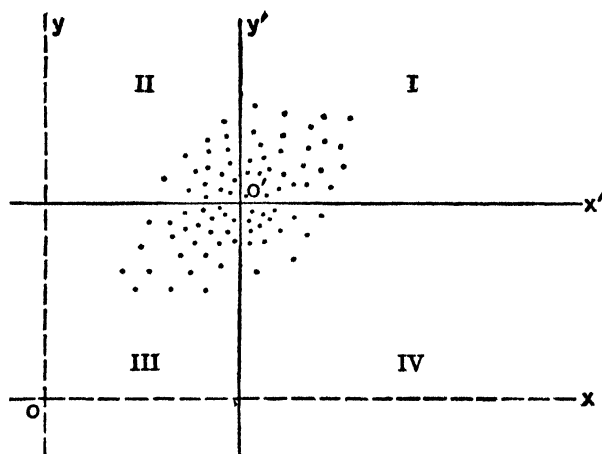


FIG. 62

In Fig. 62 the origin of the $x'y'$ -axes is taken at (\bar{x}, \bar{y}) . Then the points of the scatter diagram are distributed over the four quadrants of the $x'y'$ -plane. The coordinates of the points in the four quadrants have algebraic signs as follows. In quadrant

- I, x' and y' are positive;
- II, x' is negative and y' is positive;
- III, x' and y' are negative;
- IV, x' is positive and y' is negative.

Therefore, the product $x'y'$ is positive for all dots which occur in quadrants I and III and negative for all dots in quadrants II and IV. The algebraic sum of all such products describes the distribution of the dots over the quadrants. When this sum is positive the trend of the dots is through quadrants III and I; when it is negative the trend is through II and IV; and when zero there is no trend, the dots being equally distributed over the four quadrants in the sense that the positive products of $x'y'$ balance the negative products. Consequently, a natural measure of correlation for the sample would be obtained by summing the products $x'y'$ for all the observed values and taking the average by dividing the result by N . Moreover, if we first express x' and y' in units of their respective standard deviations we obtain a measure of correlation which is independent of the original units. This measure is

universally denoted by r . Thus we have in symbols,

$$\begin{aligned}
 (15.16) \quad r &= \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \\
 &= \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)
 \end{aligned}$$

It is called the *product-moment coefficient of correlation* or *Pearson's correlation coefficient*.

From the definition of covariance, (15.8), it is clear that

$$\begin{aligned}
 (15.17) \quad r &= s_{xy}/s_x s_y \\
 &= \text{cov}(x, y) / [\text{var}(x) \cdot \text{var}(y)]^{1/2}
 \end{aligned}$$

If the standardized variate $(x - \bar{x})/s_x$ is denoted by z and the standardized variate $(y - \bar{y})/s_y$ by w , then r is simply the covariance of z and w , that is, the arithmetic mean of the products z, w .

For the purpose of calculation with ungrouped variates, the most convenient formula for r is

$$(15.18) \quad r = \frac{N \sum xy - \sum x \sum y}{[N \sum x^2 - (\sum x)^2]^{1/2} [N \sum y^2 - (\sum y)^2]^{1/2}}$$

The calculation may frequently be simplified by making use of the following theorem.

Theorem 1. *The value of r is independent of the origin of reference and the units of measurement.*

Proof: Let

$$u = \frac{x - x_0}{h}, \quad v = \frac{y - y_0}{k}$$

Then

$$x = uh + x_0, \quad y = vk + y_0, \quad s_x = h s_u, \quad s_y = k s_v$$

Since $x - \bar{x} = (u - \bar{u})h$ and $y - \bar{y} = (v - \bar{v})k$, we have, on substituting in (15.16),

$$\begin{aligned}
 (15.19) \quad r &= \frac{1}{N} \sum_i \left(\frac{u_i - \bar{u}}{s_u} \right) \left(\frac{v_i - \bar{v}}{s_v} \right) \\
 &= \frac{N \sum uv - \sum u \sum v}{[N \sum u^2 - (\sum u)^2]^{1/2} [N \sum v^2 - (\sum v)^2]^{1/2}}
 \end{aligned}$$

which is exactly the same formula as (15.18), with u and v instead of x and y .

Example 3. To illustrate the formulas we will compute the value of r for the following data. Here x = Brokers' Loans in billions of dollars and y = *The Annalist's* index of the prices of fifty rail and industrial stocks. We choose $u = x - 5.00$ and $v = y - 250$, $h = k = 1$.

Month	x	y	u	v	uv	u^2	v^2
J	5.33	248	0.33	-2	-0.66	0.1089	4
F	5.67	248	.67	-2	-1.34	.4489	4
M	5.65	243	.65	-7	-4.55	.4225	49
A	5.56	249	.56	-1	-.56	.3136	1
My	5.53	235	.53	-15	-7.95	.2809	225
J	5.28	265	.28	15	4.20	.0784	225
Jy	5.77	282	.77	32	24.64	.5929	1024
A	6.02	303	1.02	53	54.06	1.0404	2809
S	6.35	290	1.35	40	54.00	1.8225	1600
O	6.80	230	1.80	-20	-36.00	3.2400	400
N	4.88	201	-.12	-49	5.88	.0144	2401
D	3.45	206	-1.55	-44	68.20	2.4025	1936
Sums			6.29	0	159.92	10.7659	10678

Computations:

$$12\sum u^2 - (\sum u)^2 = 89.6267$$

$$12\sum v^2 - (\sum v)^2 = 128136$$

$$12\sum uv - (\sum u)(\sum v) = 1919.04$$

$$r = \frac{1919.04}{(89.6267 \times 128136)^{1/2}} = \frac{1919.04}{3388.9} = 0.57$$

In large-scale computations the use of a calculating machine is almost essential. Students interested in such work should consult Reference 1.

A subscript notation is attached to r when there are several variates, thus, r_{xy} for the correlation between x and y , r_{xz} for that between x and z , r_{12} for the correlation between x_1 and x_2 , etc.

15.4 Relation between Coefficients of Regression and Correlation. From equations (15.9), (15.14), and (15.17) we see that

$$(15.20) \quad r^2 = bb'$$

but in computing r from this relation we must give it the sign of b and b' (both of which have the same sign as s_{xy}). For the data of Example 2, illustrated in Fig. 61, $b = -0.641$, $b' = -0.408$, $r = -(0.261)^{1/2} = -0.51$. The following quotation from Snedecor (Reference 2) throws light on the distinction between regression and correlation:

In other words, r is the geometric mean of the two regression coefficients. . . . This serves to clarify the relation of the two coefficients, correlation and regression, in measuring relationship. The latter is the appropriate one if one variable, y , may be designated as dependent on the other, x . Values of y may be partly controlled or caused by x , as when the available amounts of some glandular secretion cause differences in the sizes of organisms. Or, y may be subsequent to x , as weight gain in nutrition experiments follows the measurement of initial weight. In such cases, the regression of y on x is usually the statistic that furnishes the information desired. It is then appropriate to attempt to estimate the value of y from a knowledge of the corresponding value of x . Correlation, on the other hand, is the appropriate measure of the relation between two variates like statures of sister and brother. The two heights are known to be associated through the complex mechanism of inheritance, but neither may be looked upon as a consequence of the other. In this sense correlation is a two-way average of relationship, while regression is directional. Of course, there are many variates whose relationship may be studied by means of either correlation or regression, or both. It is necessary only to keep clearly in mind the character of the relation being considered.

The two regression lines may be written

$$(15.21) \quad Y - \bar{y} = r \frac{s_y}{s_x} (x - \bar{x})$$

$$(15.22) \quad X - \bar{x} = r \frac{s_x}{s_y} (y - \bar{y})$$

or, in terms of the standardized variates, z and w ,

$$(15.23) \quad W = rz$$

$$(15.24) \quad Z = rw$$

The coefficient of correlation is therefore the slope of the regression line of w on z , and its reciprocal is the slope of the regression line of z on w (with equal

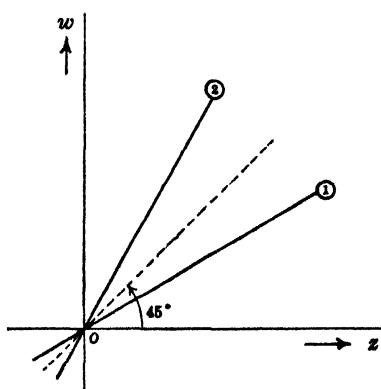


FIG. 63. REGRESSION LINES IN STANDARDIZED VARIATES

scales for z and w , the second regression line makes with the w axis the same angle that the first regression line makes with the z axis, namely, the angle whose tangent is r). Both lines go through the origin (see Fig. 63, which is drawn for the case $r > 0$).

The larger the value of r , numerically, the smaller the angle between the two regression lines, and the narrower the band of dots in the scatter diagram. As we shall see, r is always between -1 and 1 . When $r = 1$ or -1 , the two regression lines coincide. When $r = 0$, they cut at right angles.

15.5 Interpretation of the Coefficient of Correlation. It must be emphasized that only when the trend of dots in the scatter diagram is a straight line

can r be regarded as a useful measure of the degree of association between x and y . If the trend is straight, a value of r near 0 means very little association between x and y , that is, x and y are practically independent variates, and if r is near 1 or -1 , y is highly dependent on x (or x on y). However, if the trend line is curved, it is possible for r to be very nearly zero and yet y to be highly dependent on x , as in Fig. 64 (see Reference 3). A different measure of association is appropriate in such cases and will be discussed in Chapter XVI.

If x and y are *independent*, the coefficient of correlation r is zero, but if $r = 0$, x and y are not necessarily independent. They are merely *uncorrelated*. Incidentally, the phrase "independent variables" in the statistical sense should not be confused with the phrase "independent variables" which is used in the ordinary sense of analysis to designate the variables on which a specified function depends. However, the two usages, though quite distinct, are not fundamentally contradictory, since functional dependence can be regarded as a limiting case of statistical dependence.

The data should be reasonably homogeneous. If the dots in the scatter diagram show a tendency to cluster in two or more groups, a spuriously high value of r may result, due merely to the heterogeneity of the data. Thus in Fig. 65, the correlation for the two groups of values taken together would be quite high, whereas each group alone would give a correlation coefficient near zero. If, on examining the data, some reasonable basis can be found for separating these two groups, it is probably best to compute a separate coefficient for each.

An observed positive (or negative) value of r in a sample is not necessarily an indication that the true correlation in the population is different from zero. We shall discuss the question of the significance of r in a later section and merely remark here that the smaller the sample, the less the significance of a given value of r . For a class of 19 students, a coefficient of 0.34 was found between their final marks in mathematics

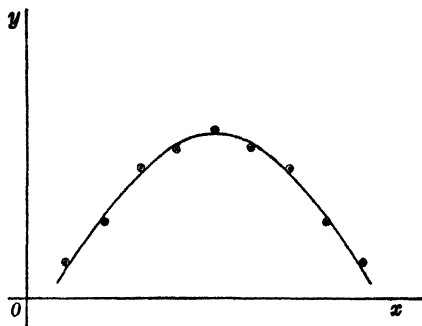
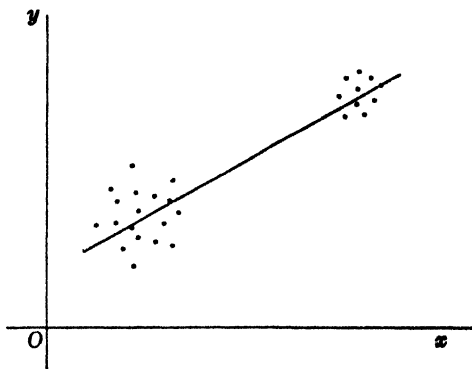
FIG. 64. $r = 0$ 

FIG. 65

and the initial letters of their surnames (A being given the value 1, B the value 2, and so on). Since it seems very unlikely that these variates are other than independent, the observed r must be purely a sampling fluctuation from a true value of zero.

Even if x and y are independent of each other, it may be that both vary roughly in the same way (or in opposite ways) with time, and hence, if observations are spread over a long interval of time, a spurious correlation may be found. Yule observed a very high correlation ($r = 0.951$) between the proportion of marriages celebrated in Anglican churches (in England and Wales) from 1866 to 1911 and the standardized mortality rate over the same period. The reason is that both these variables show a well-marked downward trend over the period. If the trend is eliminated and only the residuals are correlated we find, as we should expect, a value which is too low to be significant (0.19 in fact). Yule called correlations such as this *nonsense correlations*. Another example was pointed out by the Norwegian statistician L. V. Charlier, who found a correlation of 0.86 between the size of the stork population in Oslo over a period of about 40 years and the number of babies born there each year.

When there is a real correlation between two variables, it may be that a change in the one variable is the *cause* of a change in the other. If so, the former would be taken as x and the latter as y , and we should be interested in the regression of y on x . Sometimes, however, it is not clear which variable corresponds to cause and which to effect. In economics, an increase in price may stimulate an increase in production, but also an increase in production may be the cause of a lowering of price; the nature of the particular commodity and its market must be considered.

15.6 Variation Around the Regression Line. The average concentration of the points in the scatter diagram around the regression line of y on x may be measured by the expression $(\sum d_i^2)/N$, where d_i is the difference between an observed y_i and the corresponding Y_i calculated from the regression line. This expression may be regarded as the variance of the y values in the sample around the regression line, and we shall denote it by s_{ey}^2 , which is usually called the *variance of estimate*. From the definition of d_i ,

$$\begin{aligned}Ns_{ey}^2 &= \sum d_i^2 = \sum (y_i - Y_i)^2 \\&= \sum [y_i - \bar{y} - b(x_i - \bar{x})]^2 \\&= \sum (y_i - \bar{y})^2 + b^2 \sum (x_i - \bar{x})^2 - 2b \sum (x_i - \bar{x})(y_i - \bar{y}) \\&= Ns_y^2 + b^2 Ns_x^2 - 2bNs_{xy}\end{aligned}$$

Now, by (15.9) and (15.17),

$$(15.25) \quad b^2 s_x^2 = b s_{xy} = r^2 s_y^2$$

so that

$$(15.26) \quad s_{e,y}^2 = (1 - r^2)s_y^2$$

Since $s_{e,y}^2$ cannot be negative, it is clear that, as stated previously, r^2 cannot be greater than 1, or, in other words, $-1 \leq r \leq 1$.

If the deviations d_i of the dots from the regression line are expressed in *standard units*,

$$\begin{aligned} \sum d_i^2 &= N s_{e,y}^2 / s_y^2 \\ &= N(1 - r^2) \end{aligned}$$

or

$$(15.27) \quad r^2 = 1 - \sum d_i^2 / N$$

We can consider $1 - \sum d_i^2 / N$ as a measure of the goodness of fit of the regression line (15.23) to the points of the scatter diagram, expressed in standard units, so that the greater the numerical value of r the better the fit.

Since Y_i is the estimated value given by the regression line for a given x_i , that is,

$$Y_i = a + bx_i$$

we have

$$\sum Y_i = Na + b \sum x_i = \sum y_i$$

by the first equation of (15.2). This implies, on dividing by N , that $\bar{Y} = \bar{y}$, so that the mean of the Y_i is the same as the mean of the y_i . We now prove the following theorem.

Theorem 2. *The variance of the Y_i is r^2 times the variance of the y_i .*

Proof:

$$\begin{aligned} N s_Y^2 &= \sum (Y_i - \bar{Y})^2 = \sum (Y_i - \bar{y})^2 \\ &= \sum b^2 (x_i - \bar{x})^2, \text{ by (15.9) and (15.10)} \\ &= N b^2 s_x^2 \\ &= N r^2 s_y^2, \text{ by (15.25),} \end{aligned}$$

so that

$$(15.28) \quad r^2 = s_Y^2 / s_y^2$$

This means that r^2 is the ratio of the variance of the computed Y_i to the variance of the observed y_i . The variance of the Y_i is sometimes called the *explained variance* — it is that part of the total variance which is accounted for or explained by the regression of y on x . From (15.26),

$$(15.29) \quad s_{e,y}^2 = s_y^2 - s_Y^2$$

and so is equal to the *residual* or *unexplained* variance.

Theorem 3. *The correlation between the y , and the Y , is the same as that between the x , and the y .*

Proof:

$$\begin{aligned} r_{Yy} &= \frac{\frac{1}{N} \sum (Y - \bar{Y})(y - \bar{y})}{s_Y s_y} \\ &= \frac{\frac{1}{N} \sum b(x - \bar{x})(y - \bar{y})}{bs_x s_y}, \text{ since } \bar{Y} = \bar{y} \\ &= r_{xy} \end{aligned}$$

Theorem 4. $\sum d^2 = \sum y^2 - a \sum y - b \sum xy$

Proof:

$$\begin{aligned} \sum d^2 &= \sum (y - a - bx)(y - a - bx) \\ &= \sum y(y - a - bx) - a \sum (y - a - bx) - b \sum x(y - a - bx) \end{aligned}$$

But, by (14.8) and (14.9),

$$\sum (y - a - bx) = 0, \quad \sum x(y - a - bx) = 0$$

so that

$$\begin{aligned} \sum d^2 &= \sum y(y - a - bx) \\ &= \sum y^2 - a \sum y - b \sum xy \end{aligned}$$

This theorem provides a check on the calculation of $\sum d^2$. It is usually necessary, however, to know a and b accurately to several significant figures, because the factors $\sum y$ and $\sum xy$ may be relatively large while the right-hand side as a whole is quite small.

All the results in this section may be applied, if desired, to the regression line of x on y . It is merely necessary to interchange x and y and write b' for b .

If the foregoing theorems are intended to apply to a *population* of pairs of values (x, y) , from which the observed N pairs form a random sample, a slight modification is necessary. Just as the sample variance s_y^2 is not an unbiased estimate of the population variance σ_y^2 , so the variance of estimate s_{xy}^2 is not an unbiased estimate of the population variance of estimate σ_{xy}^2 . It turns out that $Ns_y^2/(N-1)$ and $Ns_{xy}^2/(N-2)$ are unbiased estimates of σ_y^2 and σ_{xy}^2 , respectively. If we denote these estimates by $\hat{\sigma}_y^2$ and $\hat{\sigma}_{xy}^2$ we have, instead of (15.26), the relation

$$(15.30) \quad (N-2)\hat{\sigma}_{xy}^2 = (1-r^2)(N-1)\hat{\sigma}_y^2$$

15.7 Significance of the Regression Coefficients. The regression coefficient b , which is the slope of the trend line of y on x , is usually determined from a sample of N pairs of values of x and y . The sample regression is not usually of great interest in itself, but it enables us to form an estimate of the true regression coefficient β in the population from which the sample is taken. If we assume (1) that the true regression is linear, given by $\eta = \alpha + \beta x$; (2) that each y_i in the population is a value of a random normal variate, independent of the other y_i 's; (3) that the mean of the y , in the population for a given x , is the corresponding value of η , $\eta_i = \alpha + \beta x_i$; and (4) that the variance of the y , is the same for all values of x , namely, σ_{ey}^2 ; then we can prove that b is normally distributed about β with variance σ_{ey}^2/Ns_x^2 . Since, however, we do not know σ_{ey}^2 , but only the estimate $\hat{\sigma}_{ey}^2$, we substitute this for σ_{ey}^2 and can then show that $N^{1/2}s_x(b - \beta)/\hat{\sigma}_{ey}$ has Student's t distribution with $N - 2$ degrees of freedom. By (15.30)

$$\hat{\sigma}_{ey} = N^{1/2}s_y(1 - r^2)^{1/2}/(N - 2)^{1/2}$$

so that

$$\begin{aligned} (15.31) \quad t &= (b - \beta) \frac{s_x}{s_y} \left(\frac{N - 2}{1 - r^2} \right)^{1/2} \\ &= (b - \beta) \frac{r}{b} \left(\frac{N - 2}{1 - r^2} \right)^{1/2} \end{aligned}$$

In order to determine whether an observed value of b differs significantly from zero, we put $\beta = 0$ in (15.31) and calculate

$$(15.32) \quad t = r \left(\frac{N - 2}{1 - r^2} \right)^{1/2}$$

If this is greater than the value t_α corresponding to an assigned significance level and the given $N - 2$, we can say that the observed b is significant. If we merely want to know whether there is any correlation in the population, regardless of the sign, we make a two-tailed test and double the probabilities given at the head of Table II in the Appendix.

Alternatively, we can use (15.31) to establish confidence limits for β . Thus if $t_{0.05}$ is the value of t , for $N - 2$ degrees of freedom, such that the probability of a numerically greater deviation is 0.05, then

$$t_{0.05} = \pm (b - \beta) \frac{r}{b} \left(\frac{N - 2}{1 - r^2} \right)^{1/2}$$

and the 95% confidence limits for β are given by

$$(15.33) \quad \beta = b \pm \frac{b}{r} t_{0.05} \left(\frac{1 - r^2}{N - 2} \right)^{1/2}$$

Example 4. For a sample of 10 the observed values of b and r are 0.163 and 0.582. Does the value of b differ significantly from zero?

For 8 degrees of freedom, $t_{0.05} = 2.306$. Also $b/r = 0.280$, $(1 - r^2)^{1/2}/(N - 2)^{1/2} = 0.288$, so that $\beta = 0.163 \pm 0.186$. The 95% confidence limits are -0.023 and 0.349 , and as this interval includes zero, the value of b is not significant.

The most dubious of the assumptions we have to make in getting the distribution of b is probably that of the constancy of variance of y for all values of x . A distribution of x and y values satisfying this condition is said to be *homoscedastic* (from Greek words meaning "equal scattering"). If the condition is clearly not satisfied, it is sometimes possible to transform the y variate (for example, by taking logarithms of y) so as to render it more nearly homoscedastic.

15.8 Significance of the Correlation Coefficient. We denote the true coefficient of correlation in the population by ρ (rho). If the regression is linear and $\rho = 0$ we must have $\beta = 0$, so that, as mentioned in the preceding section, the quantity $r \left(\frac{N - 2}{1 - r^2} \right)^{1/2}$ has Student's t distribution, provided y satisfies the various conditions laid down.

Example 5. From the data of Example 4, $r \left(\frac{N - 2}{1 - r^2} \right)^{1/2} = 0.582/0.288 = 2.02$, which is below the 5% significance level. The observed r is therefore not significantly different from zero. This illustrates the fact that a coefficient of correlation calculated from a sample as small as 10 is often highly unreliable.

When N is large the distribution of the sample coefficient of correlation r (if the true value ρ is zero) is nearly normal with mean zero and variance $1/(N - 1)$.

If ρ is not zero, the distribution of r is skew and mathematically complicated, even when the conditions laid down in §15.7 are satisfied. It was shown by R. A. Fisher that if we write

$$(15.34) \quad z' = \frac{1}{2} \log_e [(1 + r)/(1 - r)]$$

then z' is approximately a normal variate with mean $\zeta = \frac{1}{2} \log_e [(1 + \rho)/(1 - \rho)]$ and variance $1/(N - 3)$. The relation (15.34) can also be written $r = \tanh z'$ (hyperbolic tangent of z') and a table of this function is given in the Appendix (Table VI). With the help of this table we can easily find confidence limits for ρ from a given value of r .

Example 6. In a class of 25 students we find a correlation coefficient of 0.731 between the scores on two tests. Establish 95% confidence limits for the value of the correlation coefficient in the population.

Corresponding to $r = 0.731$, we find from the table (reading from the inside out to the margin) that $z' = 0.931$. The standard deviation of z' is $1/(22)^{1/2} = 0.213$, so that, taking

the 5% point for the normal law as 1.96, the 95% confidence limits for ζ (where $\rho = \tanh \zeta$) are given by $\zeta = 0.931 \pm 1.96(0.213) = 0.513$ and 1.349.

Consulting the table again, we see that these values correspond to $\rho = 0.472$ and 0.874, respectively, which are the required confidence limits. (ζ is the Greek letter zeta.)

F. N. David (Reference 4) has constructed diagrams, one of which is reproduced as Chart II in the Appendix, giving confidence limits for ρ corresponding to different values of r , for various sample sizes from 3 to 400. To use the diagram for the data of Example 6, we simply follow up the ordinate at the point $r = 0.731$ on the horizontal scale until we cross the two curves marked 25, one in the lower half and one in the upper half of the diagram. The corresponding values of ρ (on the vertical scale) give the confidence limits.

15.9 Accuracy of Estimate from Regression. The quantity s_{ey} , given by (15.26) is called the *standard error of estimate* because (for large samples) it is a measure of the error to be expected in estimating y for a given value of x , by means of the computed value Y . When $r = 0$, (15.21) becomes $Y = \bar{y}$ which means that the best estimate of y for *any* value of x is the mean of the y -distribution. In other words, knowledge of x is of no value in predicting y . When $r = 0$ in (15.26), $s_{ey} = s_y$. This is to be expected since the dispersion s_{ey} about the line $y = \bar{y}$ is the same as the dispersion s_y of the given y 's about their mean.

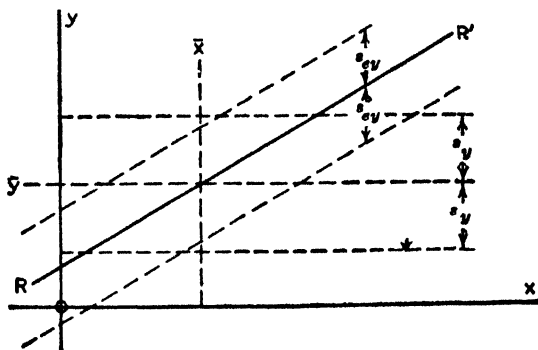


FIG. 66

In Fig. 66 parallel lines are drawn at a vertical distance of s_{ey} on either side of the regression line RR' . This strip will enclose about two-thirds of the whole distribution (assumed normal). The strip between the lines on either side of $y = \bar{y}$, at a distance s_y from it, encloses about two-thirds of the distribution when $r = 0$. As $|r|$ increases from 0, the line RR' rotates from the horizontal toward the final position it would have for $|r| = 1$, and at the same time s_{ey} decreases from s_y to 0.

As $|r|$ increases, k decreases, where $k = s_{ey}/s_y = (1 - r^2)^{1/2}$. The improvement in the estimation of y from a knowledge of the regression may be measured by $k' = 1 - k$. Table 58 gives values of k and k' for given values of r .

We see from this table that when $r = 0.5$, for example, $k' = 0.134$. This means that $s_{e,y}$ has been reduced 13.4% from the value s_y , in virtue of the correlation that exists, so that the standard error of our estimate is reduced by that much from what it would be if we used simply the estimate \bar{y} . It is clear that a high correlation is necessary in order to make a substantial reduction in the error of estimate.

TABLE 58. VALUES OF r AND THE CORRESPONDING VALUES OF k AND k'

r	k	k'
0.1	0.995	0.005
.2	.980	.020
.3	.954	.046
.4	.917	.083
.5	.866	.134
.6	.800	.200
.7	.714	.286
.8	.600	.400
.9	.436	.564
.92	.392	.608
.94	.341	.659
.96	.280	.720
.98	.198	.811
1.00	0.000	1.000

With small samples, it may be shown that the variance of y about the regression line, for any given x , is not constant, but is a function of x . It is given by

$$(15.35) \quad \text{Var } (y - Y) = \sigma_{e,y}^2 \left[1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{N s_x^2} \right]$$

and in this formula $\sigma_{e,y}^2$ must be replaced by its estimate $\hat{\sigma}_{e,y}^2 = N s_{e,y}^2 / (N - 2)$. The strip around the regression line is wider as the distance of x from \bar{x} increases. However, the variation in width is not very pronounced for moderately large values of N , and for x not too far from \bar{x} .

15.10 Calculation of r for Grouped Variates. When the sample to be studied is large, it is more convenient to replace the scatter diagram by a correlation table. We may divide the xy -plane into rectangles of convenient size, and all points of the scatter diagram falling within any rectangle are thought of as being concentrated at the center of this rectangle. A number is then written within the rectangle to designate the number of points at its center. A correlation table is therefore a two-way frequency table exhibiting the frequencies in each class interval.

Suppose Table 59 is constructed in this way for a set of average daily grades (x) and final examination grades (y) of 100 students. When the data have been thus grouped into classes, the class marks are regarded as the variate values. Thus in Table 59 there are 9 students whose daily grades are 87 and

TABLE 59

		65 — 69	70 — 74	75 — 79	80 — 84	85 — 89	90 — 94	95 — 99	
	$x \backslash y$	67	72	77	82	87	92	97	f_y
90—94	92				1	2	3	1	7
85—89	87			1	3	8	1	5	18
80—84	82	4	4	6	4	9	1		28
75—79	77	3	3	7	6	4			23
70—74	72	2	3	5	6	1	1		18
65—69	67	3	2						5
60—64	62	1							1
	f_x	13	12	19	20	24	6	6	100

whose final examination grades are 82. The last column labeled f_y represents the distribution of y variates and the last row labeled f_x represents the distribution of x variates. A correlation table is thus a bivariate distribution. In Table 59 the width of the class interval is the same for x and y , but of course this is not generally the case.

The rectangles containing the frequencies are called cells. The frequency in a typical cell is denoted by f_{xy} , meaning the frequency in the cell whose coordinates are x and y , where x and y are the mid-values of the class intervals. Both columns and rows are subdistributions of the total frequency N . Each column is a frequency distribution of y 's corresponding to a mid- x value. Similarly, each row is a frequency distribution corresponding to a mid- y value. The sum along any row is denoted by $\sum_x f_{xy}$, being the sum of the frequencies in the (x, y) cells in the x -direction. Since the marginal total for any row is the total frequency corresponding to a given value of y , it is therefore written in the column headed f_y . Thus, in Table 59, for $y = 92$,

$$\sum_x f_{xy} = 1 + 2 + 3 + 1 = 7$$

Similarly, $\sum_y f_{xy}$ denotes a summation in the y -direction of all the entries in a column, corresponding to a fixed value of x , so it denotes an entry in the

bottom row which contains the f_x frequencies. Thus, for $x = 67$

$$\sum_y f_{xy} = 4 + 3 + 2 + 3 + 1 = 13$$

The total frequency N may be obtained by adding either the marginal sub-totals f_x , the marginal sub-totals f_y , or all the cell frequencies f_{xy} . That is,

$$(15.36) \quad N = \sum_y f_y = \sum_x f_x = \sum_{xy} f_{xy}$$

The marginal sub-totals do not determine the cell frequencies uniquely. For example, we might replace the four cell frequencies in the upper right-hand corner of Table 59 by the cell frequencies shown alongside without disturbing the sub-totals.

2	2
2	4

The mean of all the x 's for such a table is given by

$$(15.37) \quad \begin{aligned} \bar{x} &= \frac{1}{N} \sum_{xy} x f_{xy} = \frac{1}{N} \sum_x x \sum_y f_{xy} \\ &= \frac{1}{N} \sum_x x f_x \end{aligned}$$

because in summing for y we keep x constant. The mean \bar{x} is the same, therefore, as the mean of the marginal distribution of x , and similarly

$$(15.38) \quad \bar{y} = \frac{1}{N} \sum_{xy} y f_{xy} = \frac{1}{N} \sum_y y f_y$$

Any column of the table is an x -array of y 's, so we shall use the symbol \bar{y}_x for the mean of a column. Similarly \bar{x}_y will denote the mean of a y -array of x 's, that is, of a row.

Theorem 5. *The mean \bar{y} is the arithmetic mean of the column means \bar{y}_x , weighted with the column frequencies f_x , that is, $\bar{y} = \frac{1}{N} \sum_x f_x \bar{y}_x$.*

Proof: By definition, $\bar{y}_x = \frac{1}{f_x} \sum_y y f_{xy}$, so that $\frac{1}{N} \sum_x f_x \bar{y}_x = \frac{1}{N} \sum_{xy} y f_{xy} = \bar{y}$.

The variances of x and y are given by

$$(15.39) \quad \begin{cases} s_x^2 = \frac{1}{N} \sum_x x^2 f_x - \bar{x}^2 \\ s_y^2 = \frac{1}{N} \sum_y y^2 f_y - \bar{y}^2 \end{cases}$$

Just as in the case of a grouped one-way frequency distribution, it was found convenient to choose an arbitrary origin and take the class interval as

unit, so we now do likewise with *both* variates. Let

$$(15.40) \quad u = (x - x_0)/h, \quad v = (y - y_0)/k$$

where h is the class-interval for x and k that for y . Then, as before,

$$(15.41) \quad \bar{x} = \bar{u}h + x_0, \quad \bar{y} = \bar{v}k + y_0$$

where

$$\bar{u} = \sum uf_u/N, \quad \bar{v} = \sum vf_v/N$$

Here f_u is the same as f_x , being the column frequency for a given x (or u). Changing the unit of x does not affect the frequencies. Similarly f_v is the same as f_y . Furthermore, as we found in Chapter VI,

$$(15.42) \quad s_x^2 = h^2 s_u^2, \quad s_y^2 = k^2 s_v^2$$

In computing r we need also the covariance of x and y , namely,

$$(15.43) \quad s_{xy} = \frac{1}{N} \sum_{xy} f_{xy} (x - \bar{x})(y - \bar{y})$$

and, from (15.40) and (15.41),

$$x - \bar{x} = h(u - \bar{u}), \quad y - \bar{y} = k(v - \bar{v})$$

so that

$$(15.44) \quad s_{xy} = \frac{hk}{N} \sum_{uv} f_{uv} (u - \bar{u})(v - \bar{v}) = hk s_{uv}$$

Then

$$(15.45) \quad r = s_{xy}/(s_x s_y) = s_{uv}/(s_u s_v)$$

so that in computing r we can work throughout in the new and simpler variates u and v . For the calculation of s_u and s_v we need to form the sums $\sum uf_u$, $\sum u^2 f_u$, $\sum vf_v$, $\sum v^2 f_v$, and a computation scheme for these is set out in Table 60. Additional columns and rows for Charlier checks are desirable, but have been omitted from the printed table for the sake of clearness. The only new point is the computation of s_{uv} .

Now

$$s_{uv} = \frac{1}{N} \sum_{uv} f_{uv} uv - \bar{u}\bar{v}$$

and

$$\sum_{uv} f_{uv} uv = \sum_u u \sum_v f_{uv} v = \sum_u u V$$

where

$$(15.46) \quad V = \sum_v vf_{v.}$$

TABLE 60. COMPUTATION OF r FOR DATA OF TABLE 59

u	-3	-2	-1	0	1	2	3	f_v	vf_v	$v^2 f_v$	U	vU	
v	x y	67	72	77	82	87	92						97
3	92				1	2	3	1	7	21	63	11	33
2	87			1	3	8	1	5	18	36	72	24	48
1	82	4	4	6	4	9	1		28	28	28	-15	-15
0	77	3	3	7	6	4			23	0	0	-18	0
-1	72	2	3	5	6	1	1		18	-18	18	-14	14
-2	67	3	2						5	-10	20	-13	26
-3	62	1							1	-3	9	-3	9
f_u		13	12	19	20	24	6	6	100	54	210	-28	(115)
uf_u		-39	-24	-19	0	24	12	18	-28				
$u^2 f_u$		117	48	19	0	24	24	54	286				
V		-7	-3	3	7	30	11	13	54				
uV		21	6	-3	0	30	22	39	(115)				

so that

$$(15.47) \quad s_{uv} = \frac{1}{N} \sum_u uV - \bar{u}\bar{v}$$

Also

$$\sum_{uv} f_{uv} uv = \sum_v v \sum_u f_{uv} u = \sum_v vU$$

where

$$(15.48) \quad U = \sum_u uf_{uv}$$

so that

$$(15.49) \quad \sum_u uV = \sum_v vU$$

The equality of these two expressions serves as a check on the calculations. Two other checks are provided by the relations

$$(15.50) \quad \sum_v U = \sum_{uv} uf_{uv} = \sum_u uf_u$$

$$(15.51) \quad \sum_u V = \sum_{uv} vf_{uv} = \sum_v vf_v$$

The U entries in the table are found by multiplying each cell frequency by the corresponding u and adding along the row. Thus for the first row, $U = 0(1) + 1(2) + 2(3) + 3(1) = 11$. The separate products may be placed in the upper right-hand corners of the cells. Similarly, the V entries are found by multiplying each cell frequency by the corresponding v and adding up the column. For the first column, $V = -3(1) - 2(3) - 1(2) + 0(3) + 1(4) = -7$. The products may be placed in the lower left-hand corners.

From the table we have

$$s_u^2 = 286/100 - (-28/100)^2 = 2.782$$

$$s_v^2 = 210/100 - (54/100)^2 = 1.808$$

$$s_{uv} = 115/100 - (-28/100)(54/100) = 1.301$$

$$r = \frac{1.301}{(2.782 \times 1.808)^{1/2}} = \frac{1.301}{2.243} = 0.58$$

Note that the data have been arranged so that the directions of increasing x and increasing y are the conventional ones along the axes Ox , Oy . A positive value of r then corresponds to a linear trend line of positive slope.

The general effect of *errors of measurement* in x and y is to decrease the coefficient of correlation, because of the increased scattering of the observations. This effect is known as *attenuation*. The effect of *grouping* is also to reduce the coefficient, because on the whole the errors caused by grouping cancel in s_{uv} but tend to increase s_u^2 and s_v^2 , unless the latter are corrected by applying Sheppard's correction. The number of cells should be large, preferably 10 or 12 each way, in order to reduce grouping errors.

Commercial charts. Computations can be expedited by the use of commercially prepared correlation charts. Several types of chart are available on the market. In her book (Reference 17, §0.4), Professor Helen M. Walker explains the merits of two of these which are recommended. She also gives the following advice to beginners: "A chart is not a crutch to help the novice. It is a means of speeding up operations after they are well understood."

15.11 Regression Lines for a Correlation Table. Suppose for each column of a correlation table we compute the mean \bar{y}_x , and place a small circle at this value of y in the middle of the column. (The easiest way to do this is to plot $\bar{v}_x = \sum v f_{uv}/f_u = V/f_u$ on the vertical v scale.) If we now fit by least squares a straight line to these column means, *each weighted with its own column frequency*, this line turns out to be the same as the ordinary regression line of y on x fitted to the scatter diagram (all the dots in a cell being taken as lying at the center of the cell). The least squares condition is

$$(15.52) \quad \sum f_x (\bar{y}_x - a_1 - b_1 x)^2 = \min$$

By partially differentiating this with respect to a_1 and b_1 and setting the results equal to zero, or (without the calculus) by treating the left-hand side as a quadratic expression in both a_1 and b_1 and using the method of §14.9, we obtain the simultaneous equations:

$$(15.53) \quad a_1 \sum f_x + b_1 \sum x f_x = \sum \bar{y}_x f_x$$

$$(15.54) \quad a_1 \sum x f_x + b_1 \sum x^2 f_x = \sum x \bar{y}_x f_x$$

Since
$$\sum_x \bar{y}_x f_x = \sum_{xy} y f_{xy} = N \bar{y}$$

and
$$\sum_x x \bar{y}_x f_x = \sum_{xy} xy f_{xy}$$

these equations become

$$(15.55) \quad a_1 + b_1 \bar{x} = \bar{y}$$

$$(15.56) \quad N a_1 \bar{x} + b_1 \sum_x x^2 f_x = \sum_{xy} xy f_{xy}$$

whence

$$(15.57) \quad b_1 = \frac{\sum xy f_{xy} - N \bar{x} \bar{y}}{\sum x^2 f_x - N \bar{x}^2} = \frac{rs_y}{s_x}$$

$$a_1 = \bar{y} - b_1 \bar{x}$$

so that the fitted line is

$$(15.58) \quad Y - \bar{y} = \frac{rs_y}{s_x} (x - \bar{x})$$

which is the regression line of y on x .

TABLE 61

	x	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5		
y	v	-4	-3	-2	-1	0	1	2	3	f_y	\bar{x}_y
125	4					2	3	2		7	47.5
115	3			1	3	1	4	4	4	17	48.1
105	2			5	7	8	11	8	7	46	45.9
95	1		2	1	10	12	9	8	2	44	44.0
85	0	1	3	12	11	7	12	7	1	54	40.7
75	-1	2	1	5	6	16	8	5		43	41.6
65	-2	2	5	5	8	8	6	1		35	38.0
55	-3	2	3	3	4	1	1			14	33.2
f_x		7	14	32	49	55	54	35	14	260	
\bar{y}_x		67.9	72.1	81.9	84.8	85.7	90.9	95.6	105.0		

Test Grades and Productive Ability

Example 7. A personnel manager in charge of hiring employees of a manufacturing plant has instituted a system of mental tests for applicants and has gathered the data shown in Table 61, where x represents the grade made on the tests and y the production ability of the applicants after they have been hired (measured as percentage of a certain standard of production).

In order to demonstrate to the company's management the connection between his mental tests and the productivity of the employees he has hired, the personnel manager does the following: (1) computes the coefficient of correlation between the two series; (2) shows what the estimated productivity of employees would be whose grades in the mental test fell on the mid-points of the class intervals of the mental test data.

The means of the columns and of the rows are given in the table. In addition, he obtains the following results:

$$\begin{aligned}\bar{x} &= 42.17, & s_y &= 17.41, & r &= 0.417, \\ \bar{y} &= 87.31, & s_x &= 8.40, & b &= r \frac{s_y}{s_x} = 0.864.\end{aligned}$$

Therefore, the line of regression of y on x is

$$Y_x - 87.31 = 0.864(x - 42.17)$$

or

$$Y_x = 0.864x + 50.88$$

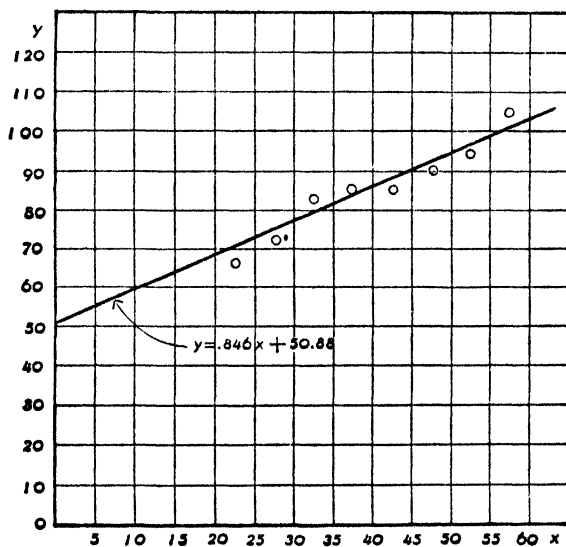


FIG. 67. MEANS OF COLUMNS AND LINE OF REGRESSION OF y ON x FOR TABLE 61

This is the equation of the line that best fits the points which designate the means of the columns (Fig. 67). Hence, for an assigned value of x , this equation gives the value of y which is the expected mean of the column defined by the assigned value of x . The personnel manager is thus prepared to predict the productivity of applicants on the basis of their mental test grades. In other words, the regression equation calculated from the records of those already hired may be used in selecting from future applicants those most likely to succeed.

The low value of $r (= 0.417)$ suggests that such prediction is not very reliable. The 95% limits for ρ are about 0.31 and 0.51, but even if we accept the observed r , the improvement in prediction due to regression is only about 9% (see Table 58, §15.9).

15.12 Variance of Estimate for a Correlation Table. We can define the variance of the dots in a column about the regression value Y_x for the center of that column by

$$(15.59) \quad s_{y \cdot x}^2 = \sum_y f_{xy} (y - Y_x)^2 / f_x$$

For example, in Table 61, for the column $u = -2$,
 $Y_x = 0.864(32.5) + 50.88 = 78.96$, $f_x = 32$

and

$$\begin{aligned} s_{y \cdot x}^2 &= [1(115 - 78.96)^2 + 5(105 - 78.96)^2 + \cdots \\ &\quad + 3(55 - 78.96)^2] / 32 = 254.7 \end{aligned}$$

We now prove the following theorem:

Theorem 6. *The arithmetic mean of the $s_{y \cdot x}^2$, each weighted with its own column frequency f_x , is equal to $s_{xy}^2 = s_y^2(1 - r^2)$.*

Proof: From (15.59)

$$\begin{aligned} \frac{1}{N} \sum_x f_x s_{y \cdot x}^2 &= \frac{1}{N} \sum_{xy} (y - Y_x)^2 f_{xy} \\ &= \frac{1}{N} \sum_{xy} f_{xy} \left[y - \bar{y} - \frac{r s_y}{s_x} (x - \bar{x}) \right]^2 \end{aligned}$$

and, as in §15.6, this reduces to $s_y^2(1 - r^2)$.

15.13 Normal Correlation Surface. A correlation table may be idealized into a surface in somewhat the same way that a histogram is idealized into a frequency curve. The concept of a surface relates to the universe from which the observed data of the table may be regarded as a sample. Let the dimensions of the cells of a table be Δx and Δy , and suppose columns are erected upon these cells with altitudes proportional to the frequencies in the cells. The result is a sort of solid histogram called a *stereogram*. Then as $\Delta x \rightarrow 0$, $\Delta y \rightarrow 0$, $N \rightarrow \infty$, the tops of the columns approach as a limit a smooth surface which is called a correlation surface. Our discussion will be confined to the case where we may assume that this limit is a normal correlation surface. In discussing this surface it is convenient to let x and y represent deviations from the respective means and to let $z = f(x, y)$ denote the frequency function representing the surface. Such a surface is shown in Fig. 68.

Any section of this surface parallel to the yz -plane is a normal curve and represents the distribution in a column at x . Similarly any section parallel to the xz -plane representing a row is a normal curve. The frequency in a

cell is measured by that portion of the volume under the surface which lies over that cell. All those cells in which the frequency is a fixed value lie on an ellipse. That is, if contour lines are drawn on the surface joining the points of equal height above the base they will be ellipses. In other words, sections of the surface parallel to the xy -plane are ellipses.

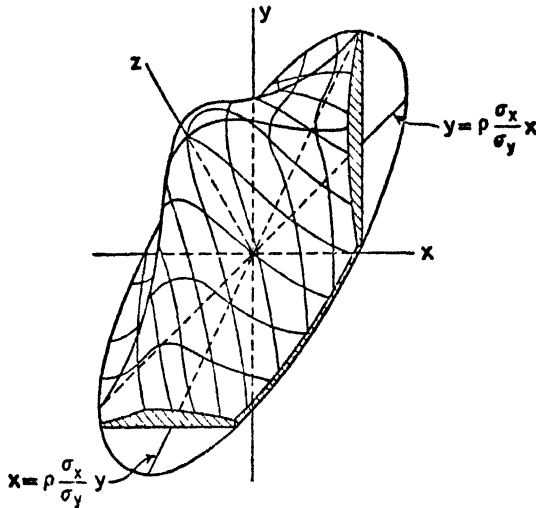


FIG. 68

We will digress here for a brief discussion of an ellipse. We may think of an ellipse as a transitional figure between a circle and a straight line, as the circle flattens out. That is to say, the limiting form of an ellipse is a circle at one extreme of the flattening process and a straight line segment at the other extreme. The degree of flatness is called the *eccentricity* of the ellipse, and it is proved in analytic geometry that the eccentricity varies from zero in the case of a circle to unity when the ellipse degenerates into a line. All ellipses having the same eccentricity whatever their size have the same relative proportions and are therefore similar in form.

The eccentricity of the elliptical contours of different normal correlation surfaces varies with the amount of correlation existing in the corresponding universe. A surface with narrow elliptical contours represents a universe in which there is high correlation, whereas if the variables are completely independent in the probability sense the contour lines are circles when the variables are expressed in standard units and when the same scale is used for both axes.

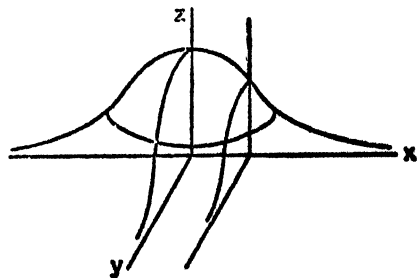


FIG. 69

If the variables are not expressed in standard units (and $\rho = 0$) then the contour lines may be ellipses, but their major and minor axes will coincide with the x - and y -axes as in Figure 69. When $\rho \neq 0$ the axes of the ellipses make an angle with the xy -axes, their major axis cuts quadrants I and III in the xy -plane if $\rho > 0$ (as in Fig. 68) and quadrants II and IV if $\rho < 0$.

The equation of a normal correlation surface is given by

$$(15.60) \quad f(x, y) = Ke^{-P}$$

where

$$P = \frac{1}{2(1 - \rho^2)} \left\{ \frac{x^2}{\sigma_x^2} - \frac{2\rho xy}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} \right\}$$

$K = 1 \div (2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2})$, and x and y represent the correlated variables referred to their respective means as origin.

By means of (15.60) an observed distribution may be fitted with the appropriate normal surface assuming that the sample might reasonably have come from such a universe. This is accomplished by replacing σ_x , σ_y , and ρ in (15.60) by the corresponding statistics calculated from the sample, multiplying by N , and taking the origin at the mean of the table.

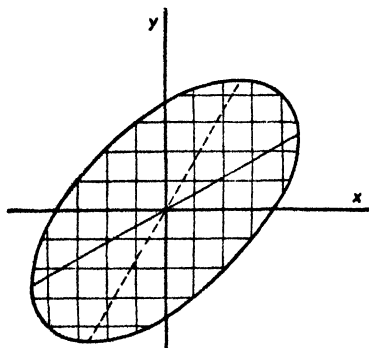


FIG. 70

Let us assume that an observed distribution has been graduated by such a surface and the theoretical cell frequencies obtained. The surface extends to infinity in the xy -plane, but contour ellipses can be obtained which will enclose any desired percentage of the given frequency when these ellipses are projected orthogonally onto the xy -plane. They are all concentric, similar, and similarly placed. Fig. 70 represents such an ellipse, say the smallest one necessary to enclose all the given cells. The systems of perpendicular chords represent the columns and rows of the table.

The graduated frequencies for each column are normal distributions whose means lie on the true regression line of y on x and whose standard deviations are in each case given by $\sigma_{ey} = \sigma_y(1 - \rho^2)^{1/2}$. To state the same thing in a slightly different way, an array of y 's corresponding to a fixed value x_1 of x is a normal distribution whose mean deviates from zero by $\rho(\sigma_y/\sigma_x)x_1$ and whose standard deviation is $\sigma_{ey} = \sigma_y(1 - \rho^2)^{1/2}$ which is independent of x_1 and therefore is the same for all such arrays. Similarly an array of x 's corresponding to a particular value y_1 of y is a normal distribution with a mean which deviates from zero by $\rho(\sigma_x/\sigma_y)y_1$, and a standard deviation of $\sigma_{ex} = \sigma_x(1 - \rho^2)^{1/2}$ which is independent of y_1 and therefore is the same for all such

arrays. A careful study of Fig. 68 will help in understanding what is meant by these statements. Cf. the four assumptions in §15.7.

For the normal correlation surface the column and row means fall exactly on the regression lines, so that the variance about the regression line is the same as the column variance, σ_{xy}^2 . This is a special case of Theorem 6, in which all the quantities being averaged are equal to each other. The distribution is homoscedastic (see §15.7).

The probability of a deviation of an observed y (for a given x) from the predicted value η (given by the true regression line) is found from the table of the normal law by expressing the deviation in standard units. If

$$(15.61) \quad z = \frac{y - \eta}{\sigma_{xy}} = \frac{y - \eta}{\sigma_y(1 - \rho^2)^{1/2}}$$

the probability of a deviation as large numerically as z is given by

$$(15.62) \quad P = \int_{-z}^z \phi(z) dz$$

In practice, η is replaced by Y , and σ_{xy}^2 by $\hat{\sigma}_{xy}^2 = Ns_{xy}^2/(N - 2)$. The distribution of z is not precisely normal, but for large N may be taken as normal with little error. For the data of Example 7, $s_{xy} = 15.8$, so that, assuming that the population is adequately represented by a normal bivariate surface, there is a probability of about 0.68 that the predicted value of $y = 79.0$ for $x = 32.5$ will not be in error by more than 16 either way.

15.14 Best-fitting Straight Line When Both Variates Are Subject to Error.

If the purpose of the regression line is to express the relationship between x and y (and not to *predict* one variate, given the other), then when y alone is subject to error we should use the regression line of y on x , and when x alone is subject to error we should use the regression line of x on y . When both variates are subject to error, neither of these lines may give the best fit.

Karl Pearson suggested the fitting of a straight line such that the sum of squares of deviations *perpendicular to the line* is a minimum. This may be satisfactory when we can use the same scale for both the variables concerned (as when correlating the stature of fathers and sons), but since in general the line depends on the choice of units for x and y it has no fundamental significance. Other methods require a prior knowledge of the standard deviations of the errors, and this knowledge is often not available. A. Wald (Reference 5) has suggested a simple method of fitting, in which the standard deviations of these errors can be estimated from the observations.

The N variates x_i, y_i are supposed to be of the form

$$(15.63) \quad x_i = \xi_i + d_i, \quad y_i = \eta_i + e_i, \quad i = 1, 2, \dots, N$$

where the true values ξ_i, η_i are connected by the linear relation

$$(15.64) \quad \eta_i = \alpha + \beta\xi_i$$

and the d_i and e_i are random variates. The d_i are uncorrelated and have a common variance σ_d^2 , the e_i are uncorrelated and have a common variance σ_e^2 , and the d_i are uncorrelated with the e_i .

For convenience we suppose that N is even ($= 2m$). The observations are divided into two equal groups, i from 1 to m , and from $m + 1$ to N , after being arranged in the order

$$x_1 \leq x_2 \leq \cdots \leq x_N$$

If the arithmetic mean of the first group is (\bar{x}_1, \bar{y}_1) and that of the second group (\bar{x}_2, \bar{y}_2) , the estimates of β and α are

$$(15.65) \quad \hat{\beta} = (\bar{y}_2 - \bar{y}_1)/(\bar{x}_2 - \bar{x}_1)$$

$$(15.66) \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

where

$$\bar{x} = (\bar{x}_1 + \bar{x}_2)/2, \quad \bar{y} = (\bar{y}_1 + \bar{y}_2)/2$$

and, furthermore, if s_x^2 , s_y^2 , s_{xy} are the sample variances and the sample covariance for x and y , estimates of σ_d^2 and σ_e^2 are

$$(15.67) \quad \hat{\sigma}_d^2 = [s_x^2 - s_{xy}/\hat{\beta}]N/(N - 1)$$

and

$$(15.68) \quad \hat{\sigma}_e^2 = [s_y^2 - \hat{\beta}s_{xy}]N/(N - 1)$$

All these estimates are *consistent*, that is to say, the probability that they differ from the true values by less than any given amount, however small, tends to one as $N \rightarrow \infty$.

The ordering of the observations by the size of the x_i is not quite independent of the errors, unless these errors are small compared with the average interval between the x_i , but in practice this arrangement is usually satisfactory.

Example 8. Applying Wald's method to the data of Example 2, §15.1, we have

$$\bar{x}_1 = (8 + 12 + 12 + 15 + 15 + 18)/6 = 13.3$$

$$\bar{x}_2 = (19 + 22 + 26 + 26 + 29 + 38)/6 = 26.7$$

$$\bar{y}_1 = (34 + 38 + 37 + 52 + 46 + 37)/6 = 40.7$$

$$\bar{y}_2 = (22 + 14 + 37 + 22 + 20 + 25)/6 = 23.3$$

so that

$$\hat{\beta} = -\frac{17.4}{13.4} = -1.30$$

$$\hat{\alpha} = 32 - 20\hat{\beta} = 58.0$$

and the line of best fit is

$$Y = 58.0 - 1.30x$$

This line is roughly midway between the two regression lines in Fig. 61.

15.15 Which Regression Should Be Used for Prediction? It does not follow that the equation which best expresses the relationship between x and y is

the best to use for predicting one variable, given a value of the other. If both x and y are subject to errors which are normally distributed, and if the true value ξ of x is also a random normal variate, then the best predicting equation for y is the ordinary regression line of y on x , even though this line is not the best expression of the relation between x and y .

There are different situations which may arise in practice. Sometimes the x and y values are measured on a *random sample* from a population and then the dots in the scatter diagram are scattered in both the x and the y directions. If so, we can safely use the regression of y on x for predicting y for a value of x to be observed in the future, and the regression of x on y for predicting x for a future value of y , whether or not the variables are subject to error.

Often, however, the x values are not random but *selected* values, and the y values are the only ones that can be regarded as random. We can still use the regression of y on x for estimating y , given x , but what are we to do about estimating x , given y ? This is a situation which arises frequently in biological experiments. For example, in assay work it may be necessary to estimate the potency of some drug preparation in terms of biological response, with the aid of a response curve based on known doses of a standard drug. The only possible regression is that of response on dosage. The other has practically no meaning.

Dealing for convenience with a large sample, we can suppose that for any assigned x the y 's are normally distributed with mean $a + bx$ and variance s_{ey}^2 . Then the probability that y lies within the limits $a + bx - 1.96 s_{ey}$ and $a + bx + 1.96 s_{ey}$ is 0.95. Of all possible pairs of (x, y) values, 95% will lie within the strip shown in Fig. 71, so that if, for a given y , we assert that the corresponding x lies in the strip — in other words, that x lies between $(y - a - 1.96 s_{ey})/b$ and $(y - a + 1.96 s_{ey})/b$ — we stand a 95% chance of being right. In the long run, if we make many such assertions for all possible values of x and y , 95% of these assertions will be true, and these limits are therefore the 95% confidence limits for x . It should be understood, however, that x is not a random variable, and for a particular y the statement about x is either true or false.

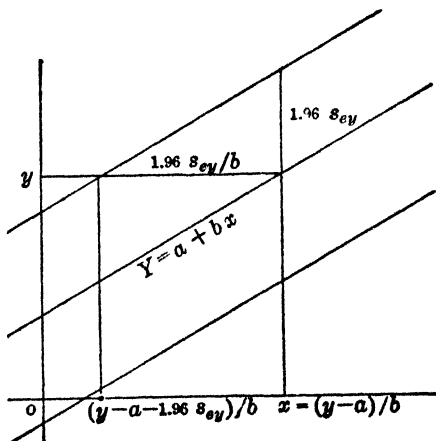


FIG. 71

A discussion by C. P. Winsor of this problem will be found in Reference 6, and more details on the modifications required when the sample is small are given by C. Eisenhart (Reference 7).

Exercises

1. Prove that (15.8) is equivalent to (15.6).
2. Write out the details of fitting the regression line of x on y by least squares, and obtaining equation (15.12).
3. The following data represent the ages of husband (x) and wife (y) for twenty couples. Calculate both regression lines and plot them on a dot diagram.

x	22	24	26	26	27	27	28	28	29	30	30	30	31	32	33	34	35	35	36	37
y	18	20	20	24	22	24	27	24	21	25	29	32	27	27	30	27	30	31	30	32

$$\text{Ans. } Y = 0.89x - 0.57.$$

$$X = 0.82y + 8.6.$$

4. Work out the estimated values Y for the given x of Exercise 3. Form the differences d_i and compute $\sum d_i$ and $\sum d_i^2$.

5. Find the value of r for the data of Exercise 3.

$$\text{Ans. } r = 0.856.$$

6. Compute s_{xy}^2 by (15.26) for the data of Exercise 3, and compare with the value $(\sum d_i^2)/20$ obtained from Exercise 4.

7. In studying a set of pairs of related variates, a statistician has completed the preliminary arithmetic and obtained the following results:

$$N = 100; \sum x^2 = 1,585,000; \sum x = 12,500; \sum xy = 1,007,425; \sum y^2 = 648,100; \sum y = 8,000. \text{ Find } \bar{x}, \bar{y}, s_x, s_y, r.$$

8. Given the following results for the heights and weights of 1000 men students:

$$\bar{y} = 68.00 \text{ in.}, \bar{x} = 150.00 \text{ lb.}, r = 0.60,$$

$$s_y = 2.50 \text{ in.}, s_x = 20.00 \text{ lb.}$$

John Doe weighs 200 lb, Richard Roe is five feet tall.

Estimate the height of Doe from his weight, and the weight of Roe from his height.

$$\text{Ans. Doe's height} = 71.75 \text{ in.}$$

$$\text{Roe's weight} = 111.6 \text{ lb}$$

9. (a) Given the following:

$$\sum x = 150,000, \sum x^2 = 22,725,000, \sum xy = 10,522,500,$$

$$\sum y = 70,000, \sum y^2 = 4,936,000, N = 1000.$$

Find \bar{x} , \bar{y} , s_x , s_y , r , and the lines of regression.

- (b) Suppose the data in (a) refer to the weight in pounds (x) and the height in inches (y) of a sample of 1000 policemen. Suppose Paul Private weighs 160 lbs and Saul Sergeant is 6 ft tall. Estimate the height of Private and the weight of Sergeant.

10. The following table contains the grades made on two tests by 25 students in mathematics. Find \bar{x} , \bar{y} , s_x , s_y , r for these data.

x	88	95	68	73	75	88	57	68	62	79	73	74	78
	80	57	65	69	74	78	72	59	47	56	67	43	
y	82	86	75	78	72	79	63	65	67	75	68	70	79
	78	51	58	65	69	68	83	80	42	43	48	47	

$$\text{Ans. } 69.8, 67.6, 12.1, 12.7, 0.79$$

11. In the following anthropometric measurements on a random sample of twenty male freshmen, taken from the Physical Education Department, x represents height, y repre-

x	y	z	x	y	z
68.5	33.6	148	65.3	33.0	136
67.2	35.0	144	65.1	34.0	144
67.7	30.2	145	64.8	37.3	170
63.8	30.0	108	69.6	33.4	154
69.9	33.0	130	68.2	31.5	122
64.7	31.0	112	68.8	32.0	141
68.4	33.0	134	72.3	35.0	159
66.4	30.2	112	67.8	33.7	134
69.1	33.3	143	71.3	31.5	136
71.0	32.3	136	63.5	33.6	126

sents chest measurement, both measurements being taken to the nearest tenth of an inch, and z represents weight to the nearest pound. Find the coefficient of correlation (a) between x and y , (b) between x and z , (c) between y and z .

12. What equation is the equivalent mathematical statement for the following words?

If the respective deviations in each series, x and y , from their means were expressed in units of standard deviations — that is, if each were divided by the standard deviation of the series to which it belongs — and plotted to a scale of standard deviations, the slope of a straight line best describing the plotted points would be the correlation coefficient r .

13. Given the standard deviations s_x , s_y for two correlated variates x and y , in a large sample:

- What is the standard error in estimating y from x if $r = 0$?
- By how much is this error reduced if r is increased to 0.25?
- How large must r be in order to reduce s_{e_y} to one-half of its value for $r = 0$?
- How large must r be to reduce s_{e_y} to one-third its value for $r = 0$?

14. Is it true that a correlation coefficient of $r = 0.6$ indicates a relationship twice as close as $r = 0.3$?

15. (For students with some knowledge of analytical geometry.) Show that the tangent of the angle between the two regression lines (15.10) and (15.12) is

$$\tan \theta = \frac{s_x s_y}{s_x^2 + s_y^2} \frac{1 - r^2}{r}$$

and between the lines (15.23) and (15.24) is $\tan \theta = (1 - r^2)/2r$. What are the values of θ for $r = 1$ and for $r = 0$?

16. The marks of a class of 12 students on a Christmas test (x) and on the final examination (y) are as follows:

Student	A	B	C	D	E	F	G	H	I	J	K	L
x	41	45	50	68	47	77	90	100	80	100	40	43
y	60	63	60	48	85	56	53	91	74	98	65	43

Estimate the final mark of a student who obtained 60 on the Christmas test but was ill at the time of the final examination. What is the standard error of this estimate?

Ans. 65, 16.

17. Specimens of similarly treated alloy steels containing various percentages of nickel are tested for toughness with the following results, where x denotes the toughness (in arbitrary units) and y the percentage of nickel:

x	47	50	52	52	54	56	58	59	60	60	62	64	65	66
y	2.5	2.7	2.8	2.8	2.9	3.2	3.2	3.3	3.4	3.5	3.5	3.6	3.7	3.8

Find the coefficient of correlation between the percentage of nickel and the toughness, as measured by the test. What is the standard error of estimate for percentage of nickel estimated from the toughness? *Ans.* $r = 0.9916$, $\hat{\sigma}_{e.y} = 0.055$.

18. The coefficient of correlation between dividends per share in 1935 and the low price of the shares during 1935 was found to be 0.817 for 64 American corporation stocks. Is this coefficient (a) significantly different from zero? (b) Significantly different from 0.75? (c) Significantly different from another sample value of 0.75 calculated from another 30 corporations?

Hint. For (b) and (c) transform to the variable z' and treat z' as normal.

19. Establish 95% confidence limits for the slope of the regression line of y on x in Exercise 3.

20. Obtain from the chart in the Appendix 95% confidence limits for ρ , in a population of married couples from which the twenty couples in Exercise 3 can be regarded as a random sample. Also transform r to the approximately normal variate z' and obtain the 95% confidence limits from the normal law.

21. In Table 59, (page 269) evaluate the following expressions:

(a) For $x = 82$,

$$\sum_y f_{xy}, \sum_y W_{xy}, \bar{y}_x$$

(b) For $y = 87$,

$$\sum_x f_{xy}, \sum_x x f_{xy}, \bar{x}_y$$

22. For the data of Table 68, (page 310) find \bar{x} , \bar{y} , s_x , s_y , and r .

Ans. $\bar{x} = 138.45$ lb, $\bar{y} = 67.82$ in., $s_x = 19.4$ lb, $s_y = 2.7$ in., $r = 0.50$.

23. For the data of Table 68, find the equations of the regression lines of x on y and of y on x .

Ans. $Y = 0.070x + 58.1$; $X = 3.50y - 99.2$.

24. Compute the value of s_{xy} for the first regression line in Exercise 23. Plot the regression line and the approximate 95% band lying along this line.

25. Referring to Exercise 8, assume that the data refer to a random sample from a normal bivariate population describing the heights and weights of senior men students in colleges and universities of the United States and Canada. Determine the probability that Doe's height is outside the interval 65.75 – 77.75 in., and the probability that Roe's weight is between 100.8 and 122.4 lb.

Ans. 0.0027; 0.5 (approx.).

26. Prove that a section parallel to the yz plane of the normal bivariate surface, with equation (15.60), is a normal curve with its mean on the regression line of y on x and with variance $\sigma_{e.y}^2 = \sigma_y^2(1 - \rho^2)$.

Hint. Write (15.60) as $f = Ke^{-P}$, where $P = (u^2 - 2\rho uv + v^2)/2(1 - \rho^2)$, $u = x/\sigma_x$, $v = y/\sigma_y$. The trace of the surface in the plane $u = u_1$ is given by putting $u = u_1$. Then $f = Ce^{-T}$, where $C = Ke^{-u_1^2/2}$ and $T = (v - \rho u_1)^2/2(1 - \rho^2) = (y - m)^2/2\sigma_{e.y}^2$, where $m = \rho x_1\sigma_y/\sigma_x$.

27. Obtain the best-fitting straight line for the data of Exercise 16 by Wald's method. Estimate by this method the standard deviations of the errors in x and y .

28. The following query and answer appeared in *Biometrics Bulletin*, vol. 1, no. 3, pp. 36-37. Investigate the references cited in the answer and justify the procedure which is recommended (under the given hypothesis).

Query. A problem that has bothered me is the fitting of regression lines when their position is restricted in some way. For example, suppose a test is made of the relationship between the number of fish present in a body of water and the average number which can be caught out of it, with a standard amount of fishing. In fitting a regression line to such data, we know that the point $(0, 0)$ must fall on the line, since if no fish are present certainly none will be caught. In other words, we have one point which is free from sampling error. The unique importance of this point will, it seems to me, make observations in its neighborhood of relatively less importance than observations at a distance from it, where there is no fixed guide-post. Do you know of any treatment of situations of this sort, by which the best straight (or curved) line could be fitted to data where there is one point which *must* be satisfied? The standard deviation from regression ("standard error of estimate") and the standard error of the regression would also be available. Or are these concepts pertinent in such a question?

Answer. Deming (§15 and §11 of reference 3, §0.4) gives both a general method and some particular solutions of your problem. Snedecor (reference 14) opens his Chapter 6 with an illustration of the simple case in which x is measured without error and the variance of y is constant for all values of x .

Observations in the neighborhood of $(0, 0)$ may or may not be of less importance than those at greater distances; it depends on the variance of y . One often finds that this variance increases with x . In fact, there are many situations in which it seems reasonable to suppose that in the sampled population the standard deviation of y is directly proportional to x . If you think this hypothesis is suitable in your fishing, the appropriate method is to calculate the ratios x/y where x is the number of fish caught and y is the total number of fish, then apply to them the statistical procedure suitable for a single variate. — George W. Snedecor.

References

1. P. S. Dwyer, "The Calculation of Correlation Coefficients from Ungrouped Data," *J. Amer. Stat. Ass.*, **35**, 1940, pp. 671-673.
2. G. W. Snedecor, *Statistical Methods* (Iowa State College Press, Ames, Iowa, 4th ed., 1946, p. 143).
3. H. L. Rietz, "On Functional Relations for which the Coefficient of Correlation is Zero," *J. Amer. Stat. Ass.*, **16**, 1919, pp. 472-476.
4. F. N. David, *Tables of the Correlation Coefficient*, (Biometrika Office, University College, London, 1938).
5. A. Wald, "The Fitting of Straight Lines if Both Variables are Subject to Error," *Ann. Math. Stat.*, **11**, 1940, pp. 284-300.
6. C. P. Winsor, "Which Regression?," *Biometrics Bulletin*, **2**, 1946, pp. 101-109.
7. C. Eisenhart, "The Interpretation of Certain Regression Methods and Their Use in Biological and Industrial Research," *Ann. Math. Stat.*, **10**, 1939, pp. 162-186.

CHAPTER XVI

FURTHER TOPICS IN CORRELATION

16.1 Reliability and Validity of Tests. In educational and psychological work, and also in the hiring of employees, considerable use is now made of various forms of mental and aptitude tests. It is desirable to know whether the results of such tests are (a) *reliable*, in the sense that a high degree of confidence can be placed in the score made by a candidate, and (b) *valid*, in the sense that an individual score on the test actually measures the ability or aptitude which it is supposed to measure. The validity of a test is estimated by the correlation of the score results with an accepted criterion of the ability in question. Thus if a test is supposed to indicate aptitude for the practice of medicine and is administered to students seeking entrance to a medical school, the validity of the test is ascertained by correlating the results for a group of students with the actual success of these students in the work of the medical school after they have been admitted.

The *reliability* of a test is judged by the correlation of the results with those of a repetition of the test, or with an alternative form of the test, on the same group of candidates. Difficulties arise in practice, however, since it cannot be assumed that a person's psychological state either remains constant in time or is unaffected by the test itself. There may be a *practice effect*, leading the candidate to do better on a second attempt, or, if the tests are too close together, there may be a *fatigue effect* leading him to do worse. Also, if two alternative forms of the test are given, it is hard to be sure that the two are of exactly the same standard of difficulty.

A method which is statistically preferable is to give one test and split it into halves, finding the coefficient of correlation r for, say, the odd-numbered and the even-numbered items on the test, care being taken to balance these items as far as practicable for length and difficulty. The reliability of the full test is measured by $(2r)/(1+r)$.

If x_i is the score made by the i th individual on one test (or half-test) and y_i is his score on the other test (or half-test), the ordinary coefficient of correlation is given by $r = s_{xy}/(s_x s_y)$, but this is not always the best estimate to use for the population coefficient of correlation. If the two scores can be supposed to have equal standard deviations σ_x and σ_y in the population, the best estimate of ρ is

$$(16.1) \quad \hat{\rho}_1 = 2s_{xy}/(s_x^2 + s_y^2)$$

and if the two scores can be supposed to have also the same means μ_x and μ_y ,

the best estimate is

$$\begin{aligned}
 (16.2) \quad \hat{\rho}_2 &= \frac{4s_{xy} - (\bar{x} - \bar{y})^2}{2(s_x^2 + s_y^2) + (\bar{x} - \bar{y})^2} \\
 &= \frac{2\sum xy - (\sum x + \sum y)^2/2N}{\sum x^2 + \sum y^2 - (\sum x + \sum y)^2/2N}
 \end{aligned}$$

The estimate (16.1) can be used when we want the correlation between a test and a re-test or between two alternative forms of the test. The estimate (16.2) is preferable for the split-half method of testing.

Example 1. In the following table x and y are the scores on two forms of the same test, made by a group of 20 pupils (d is the difference $y - x$).

TABLE 62

Pupil No.	1	2	3	4	5	6	7	8	9	10
x	9	15	10	40	19	17	18	15	24	24
y	14	22	19	37	20	34	19	20	29	24
d	5	7	9	-3	1	17	1	5	5	0
Pupil No.	11	12	13	14	15	16	17	18	19	20
x	13	13	19	16	41	35	32	20	12	17
y	28	16	28	16	46	30	41	24	11	22
d	15	3	9	0	5	-5	9	4	-1	5

We find $\bar{x} = 20.45$, $\bar{y} = 25.00$, $s_x^2 = 85.55$, $s_y^2 = 80.10$, $s_{xy} = 68.35$, $r = 0.826$. The estimate $\hat{\rho}_1$ is 0.825, so that there is very little difference between this and r . The estimate $\hat{\rho}_2$ is 0.718.

The differences of scores on the two forms of a test, or on the two halves of a split test, can be used to estimate the reliability of the test, the assumption being that the differences are due to errors of measurement. If $d = x - y$, $\sum d^2 = \sum (x^2 - 2xy + y^2) = \sum x^2 - 2\sum xy + \sum y^2$. Also $(\sum d)^2/N = [(\sum x)^2 + 2(\sum x)(\sum y) + (\sum y)^2]/N$, so that the variance of the differences d_i is given by

$$\begin{aligned}
 Ns_d^2 &= \sum d^2 - (\sum d)^2/N \\
 &= N(s_x^2 + s_y^2 - 2s_{xy})
 \end{aligned}$$

Since $s_{xy} = rs_x s_y$, this can be written

$$\begin{aligned}
 (16.3) \quad s_d^2 &= s_x^2 + s_y^2 - 2rs_x s_y \\
 &= (s_x - s_y)^2 + 2(1 - r)s_x s_y
 \end{aligned}$$

If $r = 0$ and $s_x = s_y$, $s_d^2 = 2s_x^2$, and if $r = 1$ and $s_x = s_y$, $s_d^2 = 0$. We therefore take $(\frac{1}{2}s_d^2)^{1/2}$ as a measure of the lack of reliability. In the example above,

$\sum d^2 = 993$, $\sum d = 91$, $s_d^2 = 28.95$, $s_d/\sqrt{2} = 3.80$. This may be compared with $[(s_x^2 + s_y^2)/2]^{1/2} = 9.10$, which is an average standard deviation of the scores themselves.

16.2 Analysis of Variance of Test Scores. The most satisfactory method of dealing with the question of reliability is probably to carry out an analysis of variance, which enables us to separate the parts of the variance due to individual differences between the students, to the practice effect, and to errors of measurement. The practice effect is estimated by the difference $\bar{y} - \bar{x}$, and the individual effect by the variation between the individual mean scores on the two tests.

If we regard the whole set of $2N$ scores as a single distribution, the sum of squares of deviations from the mean (called the total sum of squares) is given by

$$(16.4) \quad S_T = \sum x^2 + \sum y^2 - (\sum x + \sum y)^2/2N$$

and, when divided by the number of degrees of freedom $2N - 1$, this gives an estimate of variance of the population of scores.

The mean score of an individual on the two tests is $z = (x + y)/2$, and if the two sets of scores are independent random samples from the same population, an estimate of the population variance is given by twice the estimated variance of z . The sum of squares for individuals is, therefore,

$$(16.5) \quad S_i = 2[\sum z^2 - (\sum z)^2/N] \\ = \frac{1}{2}(\sum x^2 + \sum y^2 + 2\sum xy) - (\sum x + \sum y)^2/2N$$

and the number of degrees of freedom is $N - 1$.

On the same hypothesis as before of independent random samples, the variance of \bar{x} or \bar{y} is the population variance divided by N , so that another estimate of the population variance is provided by N times the estimate of the variance of the means. As there are only two means, their estimated variance is $[(\bar{x})^2 + (\bar{y})^2 - (\bar{x} + \bar{y})^2/2]/(2 - 1) = (\bar{y} - \bar{x})^2/2 = (\bar{d})^2/2$. The sum of squares for the practice effect is, therefore,

$$(16.6) \quad S_d = N(\bar{d})^2/2 = (\sum y - \sum x)^2/2N$$

Finally, the error variance is estimated by one-half the variance of the differences $d = y - x$, and the sum of squares for error is

$$(16.7) \quad S_e = [\sum d^2 - (\sum d)^2/N]/2 = Ns_d^2/2$$

with $N - 1$ degrees of freedom. Since

$$S_e = (\sum x^2 + \sum y^2 - 2\sum xy)/2 - (\sum y - \sum x)^2/2N$$

it is easily proved from the foregoing expressions that

$$(16.8) \quad S_r = S_e + S_d + S_o.$$

Table 63 gives the analysis of variance for the data of Example 1. On the null hypotheses that there are no real individual differences and no real practice effect, the ratio of the first and third estimates of variance in column 4, and also the ratio of the second and third estimates, are distributed as F . From Table IV of the Appendix, we see that the 5% and 1% significance levels of F for 1 and 19 degrees of freedom are 4.38 and 8.18, so that the observed ratio $207.0/15.2 = 13.6$ is highly significant. There is therefore a well-marked practice effect. The levels for 19 and 19 degrees of freedom are 2.16 and 3.03, so that the observed ratio $159/15.2 = 10.5$ is highly significant. A large value for this ratio implies a reliable test, as it indicates that differences between individuals are large compared with the error.

TABLE 63. ANALYSIS OF VARIANCE FOR SCORES ON ALTERNATIVE FORMS OF A TEST

<i>Variation Due to</i>	<i>Sum of Squares</i>	<i>Degrees of Freedom</i>	<i>Estimate of Variance</i>
Individual differences	3023.5	19	159
Practice effect	207	1	207
Error	289.5	19	15.2
Total	3520	39	90.3

For further details on the reliability of tests, the reader may consult Reference 1.

16.3 Rank Correlation. It is sometimes possible to place a group of individuals in order with respect to some characteristic without giving a definite numerical score to each individual. For instance, a judge may have to rank a group of bathing beauties for a contest, or a sales manager may rank a group of salesmen in order of efficiency. The rank is a variate which takes (except for possible ties) only the values $1, 2, \dots, N$. The mean \bar{x} is therefore $(N + 1)/2$, and the variance is given by

$$(16.9) \quad s_x^2 = \left(\sum_1^N x^2 \right) / N - \bar{x}^2 \\ = (N + 1)(2N + 1)/6 - (N + 1)^2/4 = (N^2 - 1)/12$$

Suppose now that the *same* individuals are ranked in two ways (by different judges, or on the basis of different characteristics), and that the rank of the i th individual is x_i on the first ranking and y_i on the second. If $d_i = y_i - x_i$, we have seen in §16.1 that the variance of d is given, for any pair of variates x and y , by

$$s_d^2 = s_x^2 + s_y^2 - 2rs_xs_y$$

so that

$$(16.10) \quad r = \frac{s_x^2 + s_y^2 - s_d^2}{2s_x s_y}$$

If x and y are ranks, s_x^2 and s_y^2 are both given by (16.9), and since $\bar{x} = \bar{y}$, s_d^2 is given by $(\sum d^2)/N$. On substituting in (16.10) we get

$$(16.11) \quad r = \frac{(N^2 - 1)/6 - (\sum d^2)/N}{(N^2 - 1)/6} \\ = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

This is known as *Spearman's formula* for rank correlation. It is the Pearson product-moment correlation coefficient for the ranks, treated as ordinary variates. For fairly small samples, less than 40, say, r is easier to compute by the rank method than by the exact method and thus is sometimes used even when the actual variate values are available.

Example 2. For the data of Example 1, where x and y now denote ranks on the two forms of the same test, we have

Pupil No.	1	2	3	4	5	6	7	8	9	10
x	20	14.5	19	2	8.5	11.5	10	14.5	5.5	5.5
y	19	11.5	15.5	3	13.5	4	15.5	13.5	6	9.5
d^2	1	9	12.25	1	25	56.25	30.25	1	0.25	16

Pupil No.	11	12	13	14	15	16	17	18	19	20
x	16.5	16.5	8.5	13	1	3	4	7	18	11.5
y	7.5	17.5	7.5	17.5	1	5	2	9.5	20	11.5
d^2	81	1	1	20.25	0	4	4	6.25	4	0

When, as frequently happens, there are ties in the rankings, it is customary to divide the corresponding rank numbers equally among the values concerned, using fractions where necessary. Thus if the 11th and 12th are equal they are both given the rank 11.5. However, Spearman's formula is no longer precisely equivalent to the product moment correlation coefficient, since Ns_x^2 , for example, is not equal to $N(N^2 - 1)/12$, if there are any ties in the x -ranking.

For the preceding data we find $\sum d^2 = 273.5$, $N = 20$, $r = 1 - \frac{1641}{7980} = 0.794$. The product moment coefficient for the scores themselves is 0.826.

The significance of an observed value of r may be estimated from the fact that if x and y are the ranks of *independent* random samples of size N from the same population, which need not be normal, the variance of the observed r is $1/(N - 1)$. If N is fairly large the distribution is approximately normal.

A different method of obtaining a rank correlation coefficient has been given by Kendall. (See Part Two, §8.17, and Reference 2.) This coefficient has a smaller sampling variance than Spearman's r , and its distribution tends more rapidly to the normal as N increases.

16.4 Parabolic Regression. In Chapter XIV we discussed the fitting of a straight line trend to a time series by the method of least squares, and we also dealt with certain curved trends which by a change of variable could be reduced to the linear form. Sometimes, however, we need to fit a curved trend line which cannot be so reduced, and the simplest curve is the second degree parabola.

For this curve, and in fact for any polynomial, the method of least squares gives the same result as the method of moments. If the best-fitting parabola is given by

$$(16.12) \quad Y = a + bx + cx^2$$

the statistics a , b , and c can be calculated from the equations

$$(16.13) \quad \begin{cases} \sum y = Na + \sum xb + \sum x^2c \\ \sum xy = \sum xa + \sum x^2b + \sum x^3c \\ \sum x^2y = \sum x^2a + \sum x^3b + \sum x^4c \end{cases}$$

which express the equality of the zeroth, first, and second moments of y and Y (see §14.8). These equations are obtainable also from the least squares condition

$$(16.14) \quad \sum (y - a - bx - cx^2)^2 = \min$$

by differentiating partially with respect to a , b , and c .

If the values of x are *equally spaced*, with a common interval h , and if we introduce the new variable

$$(16.15) \quad u = (x - \bar{x})/h$$

then, as in §14.11, the u are consecutive integers if N is odd, or half-integers differing by unity if N is even. In either case $\sum u^2 = N(N^2 - 1)/12 = m$, say, and $\sum u^4 = N(N^2 - 1)(3N^2 - 7)/240 = n$, say. The equations for a , b , and c are now much simplified. They are

$$(16.16) \quad \begin{cases} Na + mc = \sum y = N\bar{y} \\ mb = \sum uy \\ ma + nc = \sum u^2y \end{cases}$$

When N is even, it is convenient to double the u values so as to avoid fractions.

Example 3. Table 64 gives the number of divorces per 1000 marriages in the United States, 1900–1930. Fit a parabolic trend line to these data.

Here

$$N = 7, \quad m = 7(48)/12 = 28$$

$$n = 7(48)(140)/240 = 196$$

(The values of m and n are checked by the column totals $\sum u^2$ and $\sum u^4$). The equations for a , b , c are

$$7a + 28c = 808$$

$$28b = 465$$

$$28a + 196c = 3415$$

From the second of these, $b = 16.607$, and, from the first and third, $a = 747/7 = 106.71$, $c = 61/28 = 2.1786$. The regression equation is therefore

$$Y = 106.7 + 16.6u + 2.18u^2$$

where $u = (x - 1915)/5$. Computed values of Y are given in the last column of Table 64.

TABLE 64. DIVORCES PER 1000 MARRIAGES, U.S.A., 1900–1930

Year	y	u	uy	u^2	u^2y	u^4	Y
1900	79	-3	-237	9	711	81	76.5
1905	81	-2	-162	4	324	16	82.2
1910	88	-1	-88	1	88	1	92.3
1915	104	0	0	0	0	0	106.7
1920	134	1	134	1	134	1	125.5
1925	148	2	296	4	592	16	148.6
1930	174	3	522	9	1566	81	176.1
	808		465	28	3415	196	

Source: *Statistical Abstract of the United States*, 1951.

Fig. 72 shows the fit of the curve to the data and also illustrates very well the dangers of extrapolation. The actual values for 1935 and 1940 fall a long way below the trend line, whereas the value for 1945 is fairly close.

The geometrical meaning of the constants a , b , c is indicated in the diagram. We see that a is the ordinate at $u = 0$, b the slope of the tangent to the curve at $u = 0$, and c the difference of ordinates (at $u = 1$) between the curve and the tangent at $u = 0$. If the curve is concave upward, c is positive; if concave downward, c is negative.

16.5 Correlation Index for Non-linear Regression. We have seen in §15.6 that when the regression is linear the Pearson coefficient of correlation r is given by

$$r^2 = s_Y^2/s_y^2 = 1 - s_{ey}^2/s_y^2$$

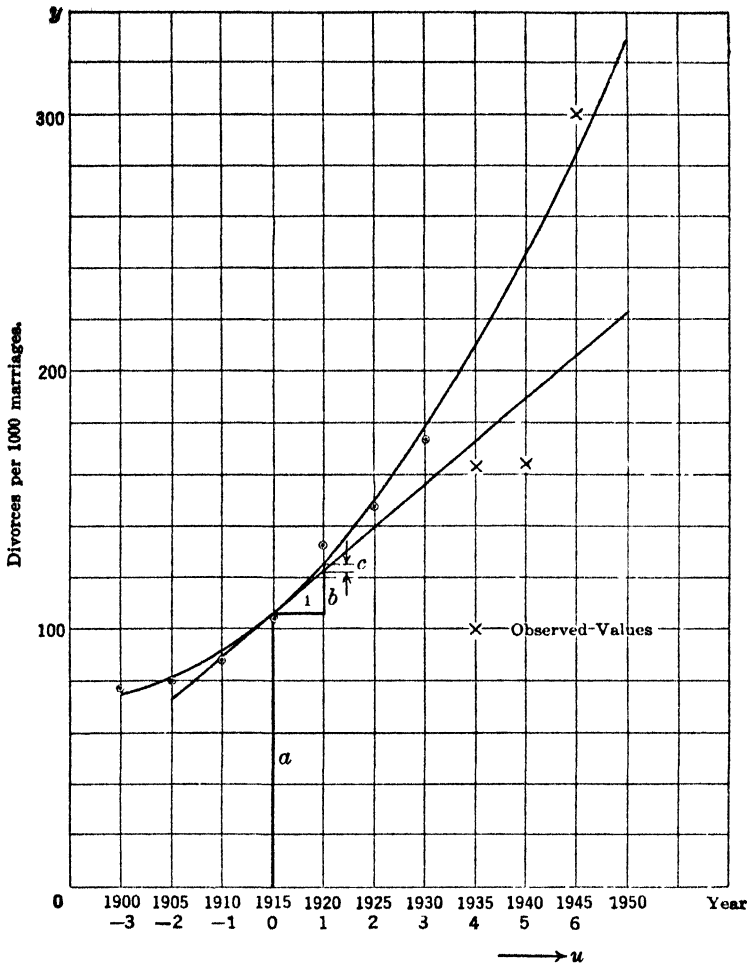


FIG. 72. PARABOLIC TREND LINE

where s_Y^2 is the variance of the computed Y 's and s_y^2 that of the observed y 's, and where $s_{e_y}^2$ is the variance of the y 's about the regression line.

When the regression is curved, we may define a *correlation index* r_c by the expression $r_c = s_{yY}/(s_Y s_Y)$ and show that

$$(16.17) \quad r_c^2 = s_Y^2/s_y^2 = 1 - s_{e_y}^2/s_y^2$$

where now the Y 's are given by the equation of the best-fitting *curved* line, and $s_{e_y}^2$ is the variance of the observed y 's about this line. The value of the correlation index depends, of course, on the particular trend line chosen, but for a given curve it is an indication of the closeness with which the observed points cluster around this line.

For a parabolic trend, the sum of squares of deviations from regression is

$$Ns_{y^2} = \sum (y - a - bu - cu^2)^2$$

and, as in the proof of Theorem 4, §15 6, we can show that

$$(16.18) \quad Ns_{y^2} = \sum y^2 - a \sum y - b \sum uy - c \sum u^2 y$$

Since $Ns_y^2 = \sum y^2 - (\sum y)^2/N$, we have from (16.17)

$$\begin{aligned} r_c^2 &= (Ns_y^2 - Ns_{y^2})/Ns_y^2 \\ &= [a \sum y - (\sum y)^2/N + b \sum uy + c \sum u^2 y]/Ns_y^2 \end{aligned}$$

On substituting for a from the first equation of (16.16) this becomes

$$(16.19) \quad r_c^2 = [b \sum uy + c(\sum u^2 y - m \sum y/N)]/Ns_y^2$$

For the data of Table 64, we have

$$Ns_y^2 = \sum y^2 - (\sum y)^2/N = 101498 - 93266 = 8232$$

$$r_c^2 = 8121/8232 = 0.9865$$

so that $r_c = 0.993$. If we fitted a straight line to the data, we should find

$$r^2 = (\sum uy)^2/mNs_y^2 = 0.9381, \quad r = 0.969$$

The fit about the parabola seems to be definitely better than that about the straight line. This is confirmed by analyzing the variance. The total sum of squares for y about \bar{y} is $Ns_y^2 = 8232$. The sum of squares about the straight regression line is $Ns_y^2(1 - r^2) = 510$ and that about the parabolic regression line is $Ns_y^2(1 - r_c^2) = 111$. The reduction due to the parabolic regression is, therefore, 399.

The significance of these sums of squares can be estimated by an analysis of variance, as in Table 65. In making an estimate of population variance about parabolic regression we divide the sum of squares by $N - 3$ (instead of by $N - 2$, as for linear regression). Three degrees of freedom are lost since three constants for the parabola are estimated from the data. The estimated variance $\hat{\sigma}_{y^2}$ is, therefore, $111/4 = 28$. Since the reduction in sum of squares due to the parabolic (over the linear) regression is 399, with 1 degree of freedom, the ratio $399/28 = 14.4$ is to be compared with the 5% and 1% values of F with 1 and 4 degrees of freedom. These values are 7.71 and 21.2 so that there is a significant reduction. That is to say, parabolic regression is definitely better than linear regression at the 5% level of significance.

TABLE 65. ANALYSIS OF VARIANCE FOR PARABOLIC REGRESSION

<i>Sum of Squares</i>	<i>Degrees of Freedom</i>	<i>Estimate of Variance</i>
Total ($N s_y^2$) = 8232	6	1372
Linear regression ($N s_Y^2$) = 7722	1	
About straight regression line } ($N s_{xy}^2$) = 510	5	102
Parabolic regression ($N s_Y^2$) = 8121	2	
About parabolic regression line } ($N s_{xy}^2$) = 111	4	28
Reduction due to parabolic regression } 399	1	399

16.6 Curves of Column and Row Means. When we are dealing with data grouped in a correlation table, the general nature of the regression may be estimated by plotting the column (or row) means and joining them by straight lines. We saw in §15.11 that the ordinary regression line of y on x

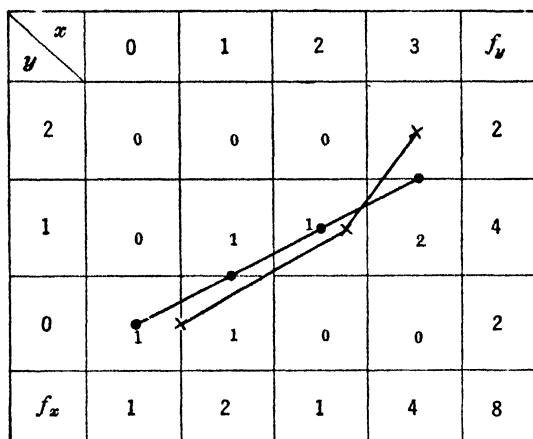


FIG. 73

gives the best-fitting *straight* line to the column means (weighted with the column frequencies), but of course the line of column means might show a well-marked departure from linearity. A simple example is provided by the table of Fig. 73, where the column means are indicated by black dots and the row means by crosses. The dots lie on a straight line, but the crosses do not.

The variance of the y values in a column about the corresponding column mean is defined by

$$(16.20) \quad s_{y,\bar{y}}^2 = \frac{1}{f_x} \sum_y f_{xy} (y - \bar{y}_x)^2$$

where $\bar{y}_x = \sum_y y f_{xy} / f_x = y_0 + kV/f_u$, in the notation of §15.10; then, by analogy with the definitions of r and r_c , we can define a *correlation ratio* E_{yx} by the relation

$$(16.21) \quad 1 - E_{yx}^2 = (\sum_x f_x s_{y,\bar{y}}^2) / N s_y^2$$

If the column means lie on a straight line, $s_{y,\bar{y}}^2$ is the same as $s_{e_y}^2$ for linear regression and then E_{yx}^2 is the same as $r^2 (= 1 - s_{e_y}^2 / s_y^2)$. In general, E_{yx}^2 is greater than r^2 , and the greater the departure of the line of column means from linearity, the greater the difference.

A similar expression for (16.21) can be written for E_{xy}^2 , where E_{xy} is the second correlation ratio, which depends upon the scatter of the observations about the line of *row means*. Although r_{xy} is always equal to r_{yx} , this symmetry is not characteristic of the correlation ratio, and in general E_{xy} is different from E_{yx} .

16.7 Calculation of Correlation Ratios for Grouped Variates. We first prove that

$$(16.22) \quad E_{yx}^2 = s_{y,\bar{y}}^2 / s_y^2$$

which may be compared with (15.28) and shows that E_{yx}^2 is the ratio of the variance of the column means to the variance of the y 's. The numerator is defined by

$$(16.23) \quad N s_{y,\bar{y}}^2 = \sum_x f_x (\bar{y}_x - \bar{y})^2$$

To prove (16.22) we need to show that

$$(16.24) \quad N s_y^2 = \sum_x f_x s_{y,\bar{y}}^2 + N s_{y,\bar{y}}^2$$

and this follows from Theorem 3 of §6.12. It is, in fact, merely equation (6.17) in a new notation. Dividing (16.24) through by $N s_y^2$, and using (16.21), we obtain

$$1 = 1 - E_{yx}^2 + s_{y,\bar{y}}^2 / s_y^2$$

which is equivalent to (16.22).

In the notation of §15.10, x and y are replaced by u and v , where $x = x_0 + hu$, $y = y_0 + kv$. Then

$$s_v^2 = k^2 s_u^2$$

and

$$Ns_y^2 = k^2 \sum_u f_u (\bar{v}_u - \bar{v})^2 = k^2 \sum_u f_u (V/f_u - \bar{v})^2$$

but

$$\sum f_u (V/f_u - \bar{v})^2 = \sum (V^2/f_u) - N\bar{v}^2$$

so that

$$(16.25) \quad E_{vz}^2 = [\sum (V^2/f_u) - N\bar{v}^2]/Ns_y^2$$

This is the formula used to calculate E_{vz}^2 . The values of V and of f_u are required in the ordinary process of finding r for a correlation table. We need merely another row along the bottom of the table (see Table 60 of §15.10) giving values of V^2/f_u . If the column means are to be plotted we need also V/f_u and a row for this can conveniently be provided first. The values of V^2/f_u are then given by multiplying V/f_u by V .

Example 3. For the data of Tables 59 and 60 (pages 269, 272) we have

f_u	13	12	19	20	24	6	6	100
V	-7	-3	3	7	30	11	13	54
$V/f_u = \bar{v}_u$	-0.54	-0.25	0.158	0.35	1.25	1.83	2.17	
V^2/f_u	3.77	0.75	0.47	2.45	37.50	20.17	28.17	93.29

$$E_{vz}^2 = [93.29 - (0.54)^2 100]/180.8 = 64.13/180.8 = 0.355$$

$$E_{vz} = 0.596$$

This is a little greater than the value of $r = 0.58$ found previously, but probably not enough to indicate a significant departure from linearity. A method of judging the significance of this difference will be given in the next section.

The formula corresponding to (16.25) for the second correlation ratio is

$$(16.26) \quad E_{zv}^2 = [\sum (U^2/f_v) - N\bar{u}^2]/Ns_u^2$$

which is the same as E_{vz}^2 with u and v (and U and V) interchanged.

For Table 60, we have

$$E_{zv}^2 = [125.11 - (0.28)^2 100]/278.2 = 0.422$$

$$E_{zv} = 0.649$$

which suggests a greater departure from linearity for the row means than for the column means.

f_v	U	U/f_v	U^2/f_v
7	11	1.57	17.29
18	24	1.33	32.00
28	-15	-0.536	8.04
23	-18	-0.783	14.09
18	-14	-0.778	10.89
5	-13	-2.60	33.80
1	-3	-3.00	9.00
100	-28		125.11

Just as the correlation coefficient r for a sample is an estimate of the true correlation coefficient ρ for the population from which the sample is taken,

so the correlation ratios are estimates of the true ratios η_{yz} and η_{zy} . Tables have been prepared from which the significance of observed values of E_{yz} or E_{zy} can be estimated. For further details and references, see Part Two (Chapter XI).

It may be well to mention that the value of E_{yz} is not independent of the classification of the data. As the class intervals become narrower, E_{yz} approaches unity. This may be understood from (16.21). If the grouping were so fine that only one item appeared in each column, then it would constitute the mean of that column. In this case $s_{y \cdot}^2$ would be zero and E_{yz} would therefore be unity. On the other hand, a very coarse grouping tends to make the value of E_{yz} approach r . Student has given a formula for *The Correction to be Made in the Correlation Ratio for Grouping* in *Biometrika*, vol. IX, pp. 316-320.

16.8 Test for Linearity of Regression with Grouped Variates. The weighted sum of squares for column means, (16.23), can be split up into a part depending on linear regression and a part depending on the deviation from linear regression, and this circumstance enables us to use the F test for the significance of the deviation from linear regression.

Let Y_x be the value of the computed y for a column, from the linear regression

$$Y_x = a + bx$$

Then

$$\bar{y}_x - \bar{y} = \bar{y}_x - Y_x + Y_x - \bar{y}$$

so that

$$(16.27) \quad \begin{aligned} \sum f_x (\bar{y}_x - \bar{y})^2 &= \sum f_x (\bar{y}_x - Y_x)^2 \\ &+ \sum f_x (Y_x - \bar{y})^2 + 2 \sum f_x (\bar{y}_x - Y_x)(Y_x - \bar{y}) \end{aligned}$$

By (16.22) and (16.23), the left-hand side of (16.27) is equal to $Ns_y^2 E_{yz}^2$. Since $\bar{y} = a + b\bar{x}$,

$$\sum f_x (Y_x - \bar{y})^2 = b^2 \sum f_x (x - \bar{x})^2 = Nb^2 s_x^2 = Ns_y^2 r^2$$

Also, the last term on the right-hand side of (16.27) vanishes, as may be shown by writing

$$\begin{aligned} &2b \sum f_x (\bar{y}_x - a - bx)(x - \bar{x}) \\ &= 2b (\sum x f_x \bar{y}_x - a \sum x f_x - b \sum x^2 f_x) \\ &- 2b \bar{x} (\sum f_x \bar{y}_x - a \sum f_x - b \sum x f_x) \\ &= 0, \text{ by (15.53) and (15.54)} \end{aligned}$$

a and b being the same as the a_1 and b_1 in those equations. Therefore we have from (16.27)

$$(16.28) \quad \begin{aligned} \sum f_x (\bar{y}_x - Y_x)^2 &= N s_y^2 E_{yx}^2 - N s_y^2 r^2 \\ &= N s_y^2 (E_{yx}^2 - r^2) \end{aligned}$$

This is the part of the sum of squares for column means which is not accounted for by linear regression. If this is large, compared with what we might expect from random sampling variation, we reject the hypothesis that the regression is really linear.

The comparison is made with the sum of squares within the columns, that is, with $\sum f_x s_{y|u}^2$ which, by (16.21), is equal to $N s_y^2 (1 - E_{yx}^2)$.

The number of degrees of freedom for variation in a column is $f_x - 1$, and, if there are p columns, the total number of degrees of freedom for the variation within columns is $\sum_x (f_x - 1) = N - p$. Also, since there are p column means fitted with a linear regression line, the number of degrees of freedom for variation from regression is $p - 2$. It^{*} can be shown that, if the parent population is uncorrelated, the ratio of $[N s_y^2 (E_{yx}^2 - r^2)] / (p - 2)$ to $[N s_y^2 (1 - E_{yx}^2)] / (N - p)$ has the F distribution with $p - 2$ and $N - p$ degrees of freedom. A significant value of F indicates a significant departure from linearity.

For the example given before,

$$E_{yx}^2 = 0.355, \quad r^2 = 0.337, \quad N = 100, \quad p = 7$$

so that

$$F = (0.018)93 / [(0.645)5] = 0.52$$

This, being less than 1, is clearly not significant. For the other line we can use the same expression for F but with E_{xy}^2 instead of E_{yx}^2 and with q (the number of rows) instead of p . For the same example, $E_{xy}^2 = 0.422$, $r^2 = 0.337$, $N = 100$, $q = 7$, so that

$$F = (0.085)93 / [(0.578)5] = 2.74$$

The 5% level for 5 and 93 degrees of freedom is about 2.31, so that there is a significant departure from linearity in the curve of row means, that is, in the regression of x on y .

16.9 Some General Remarks on Correlation. The relationship between the correlation coefficient and the correlation ratio may be clarified by Fig. 74.

For completely random scattering of the dots, with no trend, r and E are both zero (E stands for either E_{yx} or E_{xy}). If the dots lie precisely on a straight line, $r = 1$ and $E = 1$. If the dots lie on a curve as in Fig. 74 (c), such that no ordinate cuts it more than once, $E_{yx} = 1$, and if, furthermore, the dots are symmetrically placed about the y -axis, $E_{xy} = 0$ and $r = 0$.

In Fig. 74(d) the dots scatter around a definitely curved trend line, and

$$E_{yx} > r.$$

Although statistical theory gives a description of the indicated relationship between two related variables, the interpretation of the results "abounds in pitfalls easily overlooked by the unwary, while they are cantering gaily along upon their arithmetic."

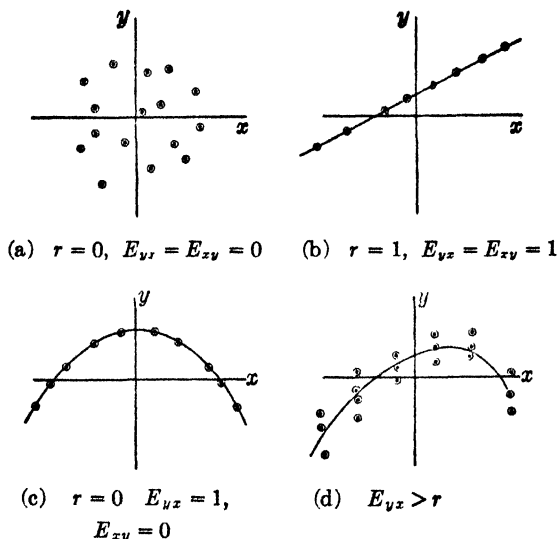


FIG. 74

The methodological side has been developed until we can find correlation coefficients by simply turning a crank, but the explanation of the meaning of the result after we find it, needs a brain . . . No amount of mathematical training and ability can take the place of the judgment and common sense that comes from a knowledge of the field in which the problem lies (Reference 3.)

In the interpretation of r one should avoid imputing any causal relationship between the variables. In this connection the following pungent remarks of Professor E. B. Wilson (Reference 4) may be appropriately quoted:

Correlation is a mutual affair between two numerical variables; the correlation coefficient r is symmetrical with respect to them. Strictly, y is not correlated with x or x with y , but x and y are correlated. Theory is very important in indicating what facts should be looked for as significant; facts are significant or important largely as they indicate theory, but neither compels the other, as the histories of theorizing and of fact finding amply demonstrate . . . Further, the value of the correlation coefficient depends on the group for which it is determined or on the universe of which that group is a fair sample. The correlation coefficient r of height and weight for a group containing humans from infancy to adult life would be different from, and in fact greater than, the coefficient for college students or for the members of a football squad; there is no such thing as the correlation coefficient *per se*.

16.10 Contingency. Frequently in biological or psychological experimental work we encounter characteristics or attributes which are not susceptible of accurate measurement, although it is possible to divide the population into two or more classes with respect to these attributes. We might, for example, divide a population into "right-handed," "left-handed," and "ambidextrous," or into "fair haired," "red-haired," "brown-haired," and "black-haired." A frequency table in which a sample from the population is classified according to two different attributes, is called a *contingency table*. It is like a correlation table, except that the different columns and rows are not assigned definite numerical values of variates x and y .

TABLE 66. CONTINGENCY TABLE

	B_1	B_2	B_3	
A_1	f_{11}	f_{12}	f_{13}	r_1
A_2	f_{21}	f_{22}	f_{23}	r_2
A_3	f_{31}	f_{32}	f_{33}	r_3
A_4	f_{41}	f_{42}	f_{43}	r_4
	c_1	c_2	c_3	N

If we have two attributes A and B and if the sample is divided into four A -categories, A_1 , A_2 , A_3 , and A_4 , and into three B -categories, B_1 , B_2 , and B_3 , we shall have a contingency table like Table 66. Here f_{ij} is the frequency of individuals in both the categories A_i and B_j , r_i is the marginal frequency of the A_i (the i th row total), c_j is the marginal frequency of the B_j (the j th column total), and N is the total number in the sample.

If the attributes A and B are independent, the probability that an individual has the attribute A_i is the same for all categories B_j and therefore is the same as the probability of A_i in the population as a whole. The *proportion* of individuals in the B_j category with attribute A_i should therefore be approximately the same for all values of j and for the right-hand margin, the differences actually observed being merely sampling fluctuations. A similar statement may be made about the proportion of individuals in the A_i category with attribute B_j . For Table 66, these statements mean that, for all i and j ,

$$f_{ij}/c_j \approx r_i/N$$

and

$$f_{ij}/r_i \approx c_j/N$$

and both are contained in the relation

$$(16.29) \quad f_{ij} \approx r_i c_j / N$$

(The symbol \approx means "approximately equal to".)

Example 4. Table 67 gives some data on hair color and eye color for 6800 males from Baden (in Germany). Is there any association between these two attributes?

TABLE 67. HAIR COLOR AND EYE COLOR

Eye Color	Hair Color				
	Fair	Red	Brown	Black	
Blue	1768	47	807	189	2811
Gray or Green	946	53	1387	746	3132
Brown	115	16	438	288	857
	2829	116	2632	1223	6800

One method of judging whether association is present is to compare the proportions, say, of black-haired people among the blue-eyed and among the brown-eyed, regarding these two classes as independent random samples, and to test whether the difference of proportions is significant according to the criterion given in Chapter XI.

These proportions are $189/2811$ and $288/857$. We can take, as an estimate of the population proportion of black-haired men, the ratio $\theta = 1223/6800$ in the whole sample. The variance of the difference of proportions in independent random samples of sizes 2811 and 857 is

$$\theta(1 - \theta) \left(\frac{1}{2811} + \frac{1}{857} \right) = 0.000225$$

The actual difference of proportions is 0.2688 and the ratio of this to its standard deviation is $0.2688/0.0150 = 17.9$. Obviously the observed difference is far too large to be accounted for as a sampling fluctuation and the conclusion is that black hair and brown eyes are definitely associated. Other proportions from the table may be treated in the same way. There is also, however, a method of testing for association in the table as a whole and this method is explained in the next section.

16.11 Chi-square Test for Association. On the assumption that the marginal distributions in Table 66 are fixed, the distribution among the cells of the table has 6 degrees of freedom. In any row there are 3 cells, but, with a fixed r_i , only 2 of these can be filled arbitrarily, the frequency in the third being then determined. Similarly with the columns, only 3 cells in each are freely adjustable. It can be proved that, under these conditions, and on the

null hypothesis that A and B are independent, the quantity

$$\chi_s^2 = \sum_{ij} (f_{ij} - \phi_{ij})^2 / \phi_{ij}$$

has approximately the χ^2 distribution* with 6 degrees of freedom, ϕ_{ij} being the expected number in the i th row and j th column, given by

$$(16.30) \quad \phi_{ij} = r_i c_j / N$$

In the general table with r rows and c columns, the number of degrees of freedom is $(r - 1)(c - 1)$.

By definition,

$$\chi_s^2 = \sum_{ij} (f_{ij})^2 / \phi_{ij} - 2 \sum_{ij} f_{ij} + \sum_{ij} \phi_{ij}$$

But

$$\sum_{ij} f_{ij} = \sum_{ij} \phi_{ij} = N,$$

so that

$$(16.31) \quad \chi_s^2 + N = \sum_{ij} (f_{ij})^2 / \phi_{ij} = N \sum_{ij} (f_{ij})^2 / r_i c_j$$

This is a convenient formula for calculating χ_s^2 . For the data of Table 67, we have

$$\begin{aligned} \frac{\chi_s^2}{N} + 1 &= \frac{(1768)^2}{2829(2811)} + \frac{(946)^2}{2829(3132)} + \cdots \\ &\quad + \frac{(288)^2}{1223(857)} \\ &= 1.158 \end{aligned}$$

so that $\chi_s^2 = 1075$. This is, of course, a very large value for 6 degrees of freedom, and the probability of obtaining as great a value on the null hypothesis is practically zero. The null hypothesis is therefore decisively rejected.

16.12 Coefficient of Contingency. Karl Pearson proposed as a measure of the association in a contingency table the coefficient

$$(16.32) \quad C = [\chi_s^2 / (\chi_s^2 + N)]^{1/2}$$

which he called a "coefficient of mean square contingency." The larger χ_s^2 , the nearer C is to 1, and the greater the degree of association. However, C can never be equal to 1 even if there is perfect association between the attributes, and the maximum value of C depends on the number of rows and columns in the table. For a 4×4 classification, for instance, the greatest possible value is 0.866. The utility of this coefficient is therefore rather doubtful. For the data of Example 4,

$$C = [1075/7875]^{1/2} = 0.37$$

* Compare equation (13.1).

16.13 The 2×2 Table. The commonest type of contingency table in practice is that with 2 rows and 2 columns, generally known as a "two-by-two table." The population is divided into two A -classes and also into two B -classes. The number of degrees of freedom, with the marginal totals fixed, is therefore only 1. An example of such a table was given in §9.2, Table 32, to illustrate relative frequencies, the A -classes being "inoculated" and "not-inoculated," and the B -classes "attacked by sickness" and "not-attacked." If the frequencies in the four cells are denoted for convenience by a, b, c, d , the value of χ_s^2 for the table is given by

$$(16.33) \quad \chi_s^2 = N(ad - bc)^2 / (r_1 r_2 c_1 c_2)$$

To prove this, we note that since the observed and expected frequencies have the same marginal totals, the difference between the observed and expected frequencies in a cell is the same for all cells, except for sign. If the expected frequencies are $\alpha, \beta, \gamma, \delta$, corresponding respectively to a, b, c, d , then $\alpha + \beta = a + b = r_1$, so that $a - \alpha = -(b - \beta)$ and similarly for the other pairs.

	A_1	A_2	
B_1	a	b	r_1
B_2	c	d	r_2
	c_1	c_2	N

Now $\alpha = r_1 c_1 / N$, so that

$$a - \alpha = a - \frac{(a + b)(a + c)}{a + b + c + d} = \frac{ad - bc}{a + b + c + d} = (ad - bc) / N$$

By definition,

$$\chi_s^2 = \frac{(a - \alpha)^2}{\alpha} + \frac{(b - \beta)^2}{\beta} + \frac{(c - \gamma)^2}{\gamma} + \frac{(d - \delta)^2}{\delta}$$

and, as we have just seen, all the numerators are equal to $(ad - bc)^2 / N^2$. Therefore

$$\begin{aligned} \chi_s^2 &= \frac{(ad - bc)^2}{N^2} \left[\frac{1}{\alpha} + \frac{1}{\beta} + \frac{1}{\gamma} + \frac{1}{\delta} \right] \\ &= \frac{(ad - bc)^2}{N} \left[\frac{1}{r_1 c_1} + \frac{1}{r_1 c_2} + \frac{1}{r_2 c_1} + \frac{1}{r_2 c_2} \right] \\ &= \frac{(ad - bc)^2}{N r_1 r_2 c_1 c_2} [r_2 c_2 + r_2 c_1 + r_1 c_2 + r_1 c_1] \\ &= \frac{(ad - bc)^2}{N r_1 r_2 c_1 c_2} [r_2 N + r_1 N] \\ &= \frac{(ad - bc)^2 N}{r_1 r_2 c_1 c_2} \end{aligned}$$

If we denote by d_1 the difference in proportions of B_1 in the categories A_1 and A_2 , and by d_2 the difference in the proportions of A_1 in the categories B_1 and B_2 , then

$$d_1 = a/c_1 - b/c_2 = \frac{a}{a+c} - \frac{b}{b+d} = \frac{ad-bc}{c_1c_2}$$

and

$$d_2 = a/r_1 - c/r_2 = \frac{a}{a+b} - \frac{c}{c+d} = \frac{ad-bc}{r_1r_2}$$

so that

$$(16.34) \quad \chi_s^2 = Nd_1d_2$$

This indicates clearly how χ_s^2 depends on the degree of association between the two attributes A and B .

For the data of Table 32, §9.2,

$$\chi_s^2 = 20(44)^2/(8)(12)(13)(7) = 4.43$$

and the probability of a value of χ^2 at least as great as this, with 1 degree of freedom, is a little less than 0.04, so that the degree of association indicated is apparently significant.

16.14 Yates' Correction. The distribution of χ^2 is continuous, whereas the distribution in a contingency table is discontinuous, the cell frequencies being necessarily integers. The approximation of the χ_s^2 distribution to a χ^2 distribution is better as N increases, but for moderate values of N the approximation is, as a rule, much improved by a correction due to Yates. This correction is analogous to that used in approximating the sum of terms of a binomial distribution by the integral of a normal curve (see §11.1), where the sum from, say, $x = a$ to $x = b$ is approximated by the integral from $a - \frac{1}{2}$ to $b + \frac{1}{2}$. In the 2×2 table the correction consists in replacing the frequency d by $d + \frac{1}{2}$ if $ad < bc$ or by $d - \frac{1}{2}$ if $ad > bc$, the remaining frequencies being adjusted accordingly, so as to keep the marginal totals unaltered. In the foregoing example, the table, corrected and rearranged, is as shown. The effect of the correction is to replace $(ad - bc)^2$ in the calculation of χ_s^2 by $(|ad - bc| - N/2)^2$, and in this case the reduction is from $(44)^2$ to $(34)^2$. (This may be checked by noting that

$$3\frac{1}{2} \cdot 2\frac{1}{2} - 4\frac{1}{2} \cdot 9\frac{1}{2} = \frac{1}{4}(35 - 171) = -34).$$

The new value of χ_s^2 is 2.65, and the probability of a value of χ^2 as large as this is 0.104.

The correction has therefore changed a probability which is significant at the 5% level to one which is non-significant.

	A_1	A_1	
B_2	$3\frac{1}{2}$	$4\frac{1}{2}$	8
B_1	$9\frac{1}{2}$	$2\frac{1}{2}$	12
	13	7	20

16.15 Fisher's Exact Method for 2×2 Tables. When the frequencies are fairly small, as in the preceding example, the probabilities of the various possible arrangements of the table, with fixed marginal totals, can be calculated exactly. The probability, in fact, of the arrangement with frequencies a, b, c, d in the four cells, is given by

$$(16.35) \quad p = (r_1!r_2!c_1!c_2!)/(a!b!c!d!N!)$$

Thus, for the data of Table 32, there are eight possibilities which may be set out as follows:

$\begin{array}{c c} 1 & 7 \\ \hline 12 & 0 \end{array}$	$\begin{array}{c c} 2 & 6 \\ \hline 11 & 1 \end{array}$	$\begin{array}{c c} 3 & 5 \\ \hline 10 & 2 \end{array}$	$\begin{array}{c c} 4 & 4 \\ \hline 9 & 3 \end{array}$
(1)	(2)	(3)	(4)
$\begin{array}{c c} 5 & 3 \\ \hline 8 & 4 \end{array}$	$\begin{array}{c c} 6 & 2 \\ \hline 7 & 5 \end{array}$	$\begin{array}{c c} 7 & 1 \\ \hline 6 & 6 \end{array}$	$\begin{array}{c c} 8 & 0 \\ \hline 5 & 7 \end{array}$
(5)	(6)	(7)	(8)

Of these, number (3) is the one actually obtained. The probabilities for the eight tables are given by

Table	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
9690 p	1	42	462	1925	3465	2772	924	99

(being all multiplied by the common factor 9690 to avoid fractions).

The probability of the arrangement (3) and of all more unlikely ones in the same direction (that is to say, with deviations from expectation in the same direction) is

$$\frac{1 + 42 + 462}{9690} = \frac{505}{9690} = 0.052$$

This corresponds to one tail of the distribution, whereas the probability calculated from χ^2 corresponds to both tails.* If the Fisher probability, as calculated previously, is multiplied by two, we get 0.104, which is very close to the probability for χ^2 as obtained after applying Yates' correction.

* By definition, the quantity χ^2 depends on the squares of the differences $a - \alpha$, etc., and so is always positive. For distributions of cell frequencies near one extreme, the values of $a - \alpha$, etc., will be opposite in sign to the corresponding values for distributions near the other extreme, but both will usually give comparatively large values for χ^2 .

If the smallest frequency in the table is *above* expectation, instead of below it, the "tail" will correspond to even higher values. Thus, for the table alongside, the expected value corresponding to the observed frequency 4 is $121/84 = 1.44$. The probability of the observed arrangement and of more unlikely ones in the same direction is, therefore, the sum of the probabilities for tables with 4, 5, 6, 7, 8, 9, 10, and 11 in the lower right-hand cell, and this is about 0.0336. The corrected value of χ^2 is 3.90, which corresponds to a two-tailed probability of 0.048. The poor agreement in this case with the exact value is not surprising in view of the very small expected frequency in one cell. With the smallest expected frequency less than 10, the exact method should as a rule be used, or alternatively, a table given in Fisher and Yates' *Statistical Tables* (Table VIII) will give an idea of the significance of the observed χ^2 .

66	7
7	4

It may be noted that when there is only one degree of freedom for χ^2 the distribution of χ (the square root of χ^2) is normal. In the example of §16.14 the value of χ^2 , after applying Yates' correction, is 2.65, corresponding to $\chi_s = 1.63$, and the probability of a value greater than 1.63 for a standard normal variate is about 0.052. The probability obtained from a table of χ^2 is double this, because if $\chi^2 > 2.65$, then either $\chi > +1.63$ or $\chi < -1.63$, and the sum of the probabilities for these alternatives is 0.104.

16.16 Problems Involving Three Variates. If we have three variates x , y , and z , which may be mutually related, the problems of correlation and regression become much more complicated, and in this book we can only touch on them very lightly. If we naturally think of z as dependent on both x and y , we can fit by least squares an equation of the form

$$(16.36) \quad Z = a + bx + cy$$

and the technique is very similar to that of fitting a parabola in the case of two variables. The equations for finding a , b , and c are

$$(16.37) \quad \begin{cases} \sum z = Na + \sum xb + \sum yc \\ \sum xz = \sum xa + \sum x^2b + \sum xyz \\ \sum yz = \sum ya + \sum xyb + \sum y^2c \end{cases}$$

The dot diagram will be a three-dimensional affair and the assumption in choosing an equation of the form (16.36) is that the dots lie more or less in a plane, scattering above and below the plane in a direction parallel to the z -axis.

There will be three ordinary correlation coefficients between the three variates, namely, r_{xy} , r_{yz} , and r_{xz} , but there are also partial and multiple correlation coefficients. The *multiple correlation coefficient* of z on x and y , denoted by $r_{z.xy}$, is the ordinary correlation coefficient between the observed

values z and the computed values Z as given by (16.36). Its square represents the part of the total variance of z which is explained by the regression of z on x and y , and it may be proved that

$$(16.38) \quad \begin{aligned} r_{z,xy}^2 &= s_z^2/s_z^2 \\ &= (r_{zx}^2 + r_{zy}^2 - 2r_{zy}r_{yz}r_{zx})/(1 - r_{zy}^2) \end{aligned}$$

The *partial correlation coefficient* of two variates x and y is defined as the ordinary correlation coefficient of x and y when the influence of the third variate z is eliminated. This influence is eliminated by subtracting from x the estimated X due to the regression of x on z , and similarly subtracting from y the estimated Y due to the regression of y on z . That is

$$(16.39) \quad \begin{cases} x_s = x - \bar{x} - r_{zx}s_z(z - \bar{z})/s_z \\ y_s = y - \bar{y} - r_{yz}s_y(z - \bar{z})/s_z \end{cases}$$

and the partial coefficient of correlation of x and y (denoted by $r_{xy.s}$) is then the ordinary coefficient for x_s and y_s . It can be proved that

$$(16.40) \quad r_{xy.s} = \frac{r_{xy} - r_{zx}r_{yz}}{[(1 - r_{zx}^2)(1 - r_{yz}^2)]^{1/2}}$$

The first step in calculating multiple and partial correlations is therefore the calculation of the ordinary correlation coefficients between each pair of variates. We shall not carry the subject any further here, and for questions of significance and the extension to more variables, we refer the student to the relevant sections of Part Two.

Exercises

1. Verify equation (16.8) from the expressions for S_T , S_x , S_y , and S_z .
2. A group of 28 students is given a one-hour test in mechanics shortly before Christmas and another similar test in February. If x and y represent scores on the two tests, $\sum x = 1092$, $\sum x^2 = 52,260$, $\sum y = 1795$, $\sum y^2 = 126,611$, $\sum xy = 75,628$. Make an analysis of variance of the data, on the lines of Table 63, §16.2, separating the variation into the part between students, the part between the two tests, and the part attributable to error. (The variation between the tests is what corresponds to the "practice effect" in Table 63. It depends on the difference of difficulty between the tests as well as on the effect of increased knowledge of the subject and practice in writing examinations.)
3. If $z = x + y$, write out a formula equivalent to (16.10) involving s_z^2 . What does this formula become when x and y are ranks?
4. Twelve salesmen are ranked in order of merit for efficiency by their manager. They are also ranked in accordance with their length of service. What indication is there of a relation between length of service and efficiency? (Garrett.)

<i>Salesmen</i>	<i>Years of Service</i>	<i>Order of Merit (Service)</i>	<i>Order of Merit (Effic.)</i>
A	5	7.5	6
B	2	11.5	12
C	10	2	1
D	8	4	9
E	6	6	8
F	4	9	5
G	12	1	2
H	2	11.5	10
I	7	5	3
J	5	7.5	7
K	9	3	4
L	3	10	11

Ans. $r = 0.80$.

5. Find Spearman's r for the following data:

	<i>Rank (x)</i>	<i>Score (x)</i>	<i>Rank (y)</i>	<i>Score (y)</i>
A	1	92	2	88
B	2	89	4	85
C	3	87	1	93
D	4	86	6	79
E	5	83	7	70
F	6	77	3	87
G	7	71	9	52
H	8	62	5	84
I	9	53	10	41
J	10	40	8	64

Ans. $r = 0.733$.

6. Calculate the Pearson coefficient of correlation for the scores in Exercise 5, and compare with the Spearman coefficient for the ranks.

7. Two judges rank seven candidates in a beauty contest as in the following table:

<i>Contestant</i>	<i>Judge 1</i>	<i>Judge 2</i>
A	2	3
B	1	4
C	4	2
D	5	5
E	3	1
F	7	6
G	6	7

Compute the correlation coefficient between the two rankings. Assuming that for a sample of 7 pairs drawn from a population of values of *independent* variates x and y , the computed rank correlation coefficient will exceed 0.714 in not more than 5% of cases and will exceed 0.893 in not more than 1%, what conclusion regarding the judges may be drawn from the above data?

8. Fit a parabola to the following data:

x	0	1	2	3	4	5	6
y	2	8	12	13	12	10	8

Calculate the correlation index r_c . *Ans.* $Y = 2.52 + 6.07x - 0.881x^2$; $r_c = 0.982$.

9. In the following table, x represents age in years for males, and y the mean vital capacity (Holzinger, *Biometrika*, 16, 1924, pp. 141-2).

x	y	x	y	x	y
19.5	227	37.5	222	55.5	201
22.5	230	40.5	218	58.5	185
25.5	230	43.5	216	61.5	200
28.5	237	46.5	210	64.5	169
31.5	227	49.5	205	67.5	160
34.5	229	52.5	193	70.5	163

Find the equation of the best-fitting parabola, and the correlation index.

Ans. $Y = 218.0 + 1.231x - 0.02935x^2$; $r_c = 0.968$.

10. Table 68 gives data on heights (y) and weights (x) of 200 freshmen. Calculate:

- (a) the two means and the two standard deviations,
 (b) the regression line of height on weight,

TABLE 68. HEIGHTS AND WEIGHTS OF 200 FRESHMEN
 (Heights to Nearest $\frac{1}{2}$ Inch; Weights to Nearest $\frac{1}{2}$ Pound)

$\begin{matrix} x \\ y \end{matrix}$	90- 99.5	100-	110-	120-	130-	140-	150-	160-	170-	180-	190-	200- 209.5	f_y
76- 77.9				1									1
74-							1	1	1	1			4
72-				1	1	1	4		1				8
70-			1	2	6	7	6	2	1	2	1	1	29
68-			2	8	17	8	9	2	1	1	1		49
66-			8	16	14	13	6	2	1			1	61
64-		3	8	7	7	3	3	1	1				33
62-	1	4	1	7	1								14
60-													0
58- 59.9		1											1
f_x	1	8	20	42	46	32	29	8	6	4	2	2	200

- (c) the regression line of weight on height,
 (d) the Pearson coefficient of correlation between height and weight,
 (e) the correlation ratios,
 (f) the column and row means.

Ans. (a) $\bar{x} = 138.45$ lb, $\bar{y} = 67.82$ in., $s_x = 19.4$ lb, $s_y = 2.74$ in.,

(b) $Y = 0.070x + 58.11$, (c) $X = 3.503y - 99.15$,

(d) $r = 0.50$, (e) $E_{yx} = 0.55$, $E_{xy} = 0.53$.

11. For Exercise 10, plot the straight regression lines and the lines of column and row means. Test the significance of the deviation from linearity in both cases.

12. Table 69 shows for male lives the correlation between the age (x) of an insured person at the time of issue of a policy and the age (y) of the insured at death. (Data of Midland Life Insurance Company, 1906–1924; see Ref. 5.)

Find E_{yx} , E_{xy} , and r , and test the significance of the departures from linearity for the curves of column means and row means.

TABLE 69. AGE OF INSURED AND AGE AT DEATH

$y \backslash x$	15	20	25	30	35	40	45	50	55	60	f_y
70									1	2	3
65								4	9	3	16
60							6	5	7	1	19
55				1		2	12	20	4		39
50					2	13	13	8			36
45				1	12	12	8				33
40			3	13	19	12					47
35		1	8	14	14						37
30		5	10	7							22
25		11	10								21
20	6	4									10
f_x	6	21	31	36	47	39	39	37	21	6	283

13. In the accompanying contingency table, x represents a rating given to each of a group of university freshmen on the basis of high school reports and y represents the final standing in degree examinations for the same group. Discuss the association between these two attributes.

$\begin{matrix} x \\ y \end{matrix}$	<i>fair</i>	<i>good</i>	<i>excellent</i>
3rd class	73	67	10
2nd class	64	84	15
1st class	5	24	28

14. In Table 70, x represents number of inches of water applied by irrigation to a crop and y represents the crop yield in bushels per acre. The numbers shown in the table are class marks of the various classes. Find the equation of regression of y on x and test whether it departs significantly from linearity.

TABLE 70. CROP YIELD AND IRRIGATION

$\begin{matrix} x \\ y \end{matrix}$	12	15	18	21	24	27	30	f_x
90						1	2	3
85					2	3		5
80			2	5	4	1		12
75		2	4	6	1			13
70			4	3	1			8
65		2		3				5
60	2			2				4
f_y	2	4	10	19	8	5	2	50

15. In a public opinion survey, the following questions were asked:

(1) Do you drink beer?

(2) Are you in favor of local option on the sale of liquor?

The results in one district were as shown in the accompanying table: Does this table provide good evidence of an association between drinking habits and opinion on the subject of local option?

	<i>Local</i>	<i>Option</i>
	<i>For</i>	<i>Against</i>
Drinkers	18	39
Non-drinkers	45	37

16. (Yule and Kendall) For a certain district in England during 20 years, records were kept of the following variables:

x = spring rainfall in inches

y = accumulated temperature in °F above 42°F in spring

z = seed-hay crop in cwt/acre

The following results were obtained:

$$\bar{x} = 4.91, \quad \bar{y} = 594, \quad \bar{z} = 28.02$$

$$s_x = 1.10, \quad s_y = 85, \quad s_z = 4.42$$

$$r_{xz} = 0.80, \quad r_{yz} = -0.40, \quad r_{xy} = -0.56$$

Calculate the regression equation of hay crop on spring rainfall and accumulated temperature.

Hint. Equation (16.36) can be written

$$Z - \bar{z} = b(x - \bar{x}) + c(y - \bar{y})$$

By solving the equations (16.37), show that

$$b = (b_{zx} - b_{yz}b_{xy})/(1 - r_{xy}^2)$$

$$c = (b_{zy} - b_{zx}b_{xy})/(1 - r_{xy}^2)$$

where $b_{zx} = s_{zx}/s_x^2$ = regression coefficient of z on x , and similarly for the other b 's.

17. Calculate the three partial correlation coefficients in Exercise 16 and also the multiple correlation coefficient of z on x and y .

18. (*Garrett*) Given that, for a group of children between the ages of 8 to 14, the ordinary coefficients of correlation between intelligence and school achievement, between intelligence and age, and between school achievement and age, are 0.80, 0.70, 0.60, respectively. What is the correlation coefficient between intelligence and school achievement in children of the same age?

Ans. 0.67

References

1. R. W. B. Jackson, *Studies on the Reliability of Tests*, (Department of Educational Research, University of Toronto, 1941).
2. M. G. Kendall, *Rank Correlation Methods*, (Griffin, 1949).
3. A. R. Crathorne, "Principles of Statistical Methodology," *J. Amer. Stat. Assoc.*, **26**, 1931, Supplement, pp. 27-32.
4. E. B. Wilson, "Correlation and Association," *J. Amer. Stat. Assoc.*, **26**, 1931, pp. 250-256.
5. "On Certain Applications of Mathematical Statistics to Actuarial Data," *The Record*, Amer. Inst. Actuaries, **13**, Part II, 1924.

REVIEW QUESTIONS AND PROBLEMS

1. Define the following terms: statistics, variate, discrete, class interval, class mark, x -array of y 's, range, regression line, sample, universe, coefficient of variation, variance.
2. Define the following terms: statistic, percentile, index number, mean absolute deviation, r th moment about the mean, standard normal variate, 95% confidence limits, 1% level of significance, null hypothesis, non-parametric statistic, rank correlation coefficient, correlation ratio.
3. Name and define five averages. Discuss their advantages and limitations.
4. What does a ratio chart show that a chart with a uniform scale does not? If you wished to plot data so as to secure the effect of a ratio chart, but had no ratio paper available, how would you accomplish the desired result?
5. Prove the following:
 - (a) The algebraic sum of the deviations of the observations from their mean is zero.
 - (b) The second moment about an arbitrary point equals the second moment about the mean increased by the square of the distance between the arbitrary point and the mean.
6. Define, and explain how to compute, the following quantities for a grouped distribution: Q_1 , Q_3 , Q , \bar{x} , s_x .

7. Give the equation of the normal curve (a) with mean μ and standard deviation σ and (b) with mean 0 and standard deviation 1.

State some of the properties of this curve.

8. Give two of the formulas for r . Discuss the use or uses of correlation in any problem that occurs to you.

9. Define the correlation ratio. Discuss its use.

10. The following is a reduced distribution of the breakfast checks at a cafeteria. Find \bar{x} and s_x .

x	f
8-12	4
13-17	8
18-22	24
23-27	21
28-32	15
33-37	14
38-42	7
43-47	4
48-52	2
53-57	1

Ans. $\bar{x} = 27.2¢$, $s_x = 9.4¢$.

11. Derive the relations which give the third and fourth moments about the mean in terms of moments about the origin. Define a_3 and a_4 . What information do they give?

12. Compute the value of a_3 and of a_4 for the distribution in Exercise 10.

13. (Walker) An algebra test was given to 400 high school children, of whom 150 were boys and 250 were girls. The results were as follows:

$n_1 = 150$	$n_2 = 250$
$\bar{x}_1 = 72.5$	$\bar{x}_2 = 73.6$
$s_1 = 7.0$	$s_2 = 6.4$

Find the mean and standard deviation of the combined groups.

14. For a normal distribution of 1500 students' grades, $\mu = 75$, $\sigma_x = 10$. What values of x will include the middle 500 grades? How many grades were below 60; above 90?

15. Suppose a distribution of 1000 breakfast checks from the cafeteria mentioned in problem 10 showed the following results: $\mu = 27¢$, $\sigma_x = 9¢$, $\alpha_3 = 0$, $\alpha_4 = 3$. On the basis of these results what is the expected frequency in the 23-27¢ class interval?

16. Given the following data as to the heights (y) and weights (x) of college men:

$$\begin{aligned} \sum y &= 6,800, & \sum y^2 &= 463,025, & \sum xy &= 1,022,250 \\ \sum x &= 15,000, & \sum x^2 &= 2,272,500, & N &= 100. \end{aligned}$$

Find \bar{x} , \bar{y} , s_x , s_y , r .

17. Derive the expression for the standard error of estimate,

$$s_{e_y} = s_y(1 - r^2)^{1/2}$$

18. Discuss the use of s_{e_y} in predictions.

19. For Table 71, (a) find the correlation coefficient, (b) find the equations of the lines of regression, (c) locate the coordinate axes through the arithmetic mean of the table and plot the lines obtained in (b).

TABLE 71. CORRELATION TABLE FOR MONTHLY RAINFALL IN INCHES AT IOWA CITY AND DES MOINES FOR 36 CONSECUTIVE YEARS

IOWA CITY

DES MOINES

$y \backslash x$	0.245	0.745	1.245	1.745	2.245	2.745	3.245	3.745	4.245	4.745	5.245	5.745	6.245	6.745	7.245	7.745	8.245	8.745	9.245	9.745	10.245	10.745	f_y
10.245																			1				1
9.745																1							1
9.245																	1	1					2
8.745													1						1				2
8.245													2		1	1							4
7.745											1											1	2
7.245									2	1		1		2	1								7
6.745			2		1		2		1		1			1			1	1					10
6.245											1												1
5.745						4	1		1				1	2			1						10
5.245			1		1	2			2		2	1		1	1				1				12
4.745					2	1	2	2	1	1	1		1			2			1				14
4.245			1		1	1	2		1	4	1	2	1										14
3.745				4	2	1	2	6	5	3	2	2	2		1								30
3.245		2	1	4	6	6	3	7	2		1						1				1		34
2.745			3	4	1	8	4	4	2	1	2			1									30
2.245			1	5	10	7	6	4	2	2													37
1.745	1	4	7	12	13	8	5	1	1	1	2		1										56
1.245	3	8	18	17	6	8	4	2		1													67
0.745	6	16	21	12	6	1	1	1			1			1									66
0.245	13	12	3	4																			32
f_x	23	42	58	62	49	47	32	27	18	15	14	7	10	5	6	5	3	2	5	0	1	1	432

20. How does the scatter diagram assist one in deciding whether the regression is linear or non-linear? Give the formulas for the correlation coefficient and for the correlation ratio of y on x , explaining the meaning of the letters used. How would you use these indices of correlation to decide whether the regression of y on x is linear or non-linear?

21. (a) In a normal distribution in which $\mu = 0$ and $\sigma_x = 4$, what proportion of the data lie where $x > 12$?

(b) If 100 of the observations lie between $x = -6$ and $x = -8$, how many of the data are there in the whole distribution?

22. (a) Expand $(a + b + c + d)^2$.

(b) The expansion of $(x_1 + x_2 + \dots + x_n)^2$ consists of the sum of the squares of the x 's plus the sum of their products taken two at a time. Express this expansion in summation notation.

23. Given N pairs of variates: $(x_{11}, x_{21}); (x_{12}, x_{22}); (x_{13}, x_{23}); \dots; (x_{1n}, x_{2n})$. Show that:

(a) the mean \bar{x} of all the variates is

$$\bar{x} = \frac{1}{2N} \sum_1^n (x_{1i} + x_{2i})$$

(b) the variance s^2 taken about the \bar{x} in (a) is

$$s^2 = \frac{1}{2N} \left[\sum_1^n (x_{1i} - \bar{x})^2 + \sum_1^n (x_{2i} - \bar{x})^2 \right]$$

Note. The quantity

$$r' = \frac{1}{Ns^2} \sum_1^n (x_{1i} - \bar{x})(x_{2i} - \bar{x})$$

where \bar{x} and s^2 are defined as in (a) and (b) is called the *intra-class* correlation coefficient. For its use see *Statistical Methods for Research Workers*, by R. A. Fisher (§38), Oliver and Boyd, London, 10th ed., 1946.

24. Let $S_r = \sum_{x=1}^N x^r$. Prove that $S_1 = N(N+1)/2$,

$$S_2 = N(N+1)(2N+1)/6, \quad S_3 = S_1^2$$

25. Sketch the graph of $y = Ae^{Bx}$, $-\infty \leq x \leq \infty$, when (a) both A and B are positive, (b) A is positive and B negative, (c) A is negative and B positive, (d) both A and B are negative.

26. For N correlated values of x and y the regression equation of y on x is found to be $y = 1 + x$. If $\bar{x} = 0$, $r = 0.5$, and $s_x = 1$, determine \bar{y} and s_y .

27. Discuss the properties of the normal correlation surface and their use in passing judgment on the reliability of predictions based upon the regression line of y on x .

28. Show how to fit a parabola by the method of moments.

29. A correlation coefficient of 0.503 is said to be highly significant. Assuming that this refers to the 1% level of significance, what is the least number of pairs of observations that must have been made in order to warrant this statement? *Ans.* 25.

Hint. Assume a normal distribution of z' , with variance $1/(N-3)$. See §15.8.

30. In Problem 13, is the difference of mean scores between the boys and the girls significant?

31. Explain the meaning of chi-square (χ^2) and how it is used as a test of the goodness of fit of a theoretical frequency curve to a distribution.

32. Describe the theoretical Poisson distribution and give examples of actual distributions which approximate to it.

33. What is the probability of getting a total of either 7 or 11 in a single throw with two dice?

34. Two groups of guinea pigs, as similar as possible, were inoculated with a certain disease. One group of 20 was used as a control. The other group of 30 was treated with a drug supposed to have curative properties. Sixteen of the control group and nine of the treated group died within a week. Discuss the significance of this result.

35. The following table gives the gains in weight (in grams) in a certain period for 10 pairs of rats, one of each pair being fed on raw peanuts and the other on roasted peanuts, the

remainder of the diet being identical for both members of the pair. Discuss whether the observed means or the variances are significantly different.

Pair No.	1	2	3	4	5	6	7	8	9	10
Raw	61	60	56	63	56	63	59	56	44	61
Roasted	55	54	47	59	51	61	57	54	62	58

36. Make an analysis of variance of the data in Problem 35.

Hint. Subtract 60 from each observation to make the numbers easier to deal with. The total number of degrees of freedom is 19, of which 9 are between pairs, 1 between diets, and 9 attributable to error. The sum of squares for the pairs may be calculated like the sum of squares for individuals in §16.2. The variations between pairs and between diets both turn out to be non-significant.

37. (*Bertrand*) The proprietor of a gambling establishment complains to the makers of a roulette wheel which he has installed that the wheel seems to favor red, and that patrons have noticed it. In 1000 trials of which he has kept a record, red has shown up 515 times, black 455 times and white 30 times, the theoretical proportions being 18 : 18 : 1. Would you consider the complaint justified?

38. The records of 1000 birth registrations in a certain area are examined and it is noted that 510 are males. What are the 95% confidence limits for the proportion of male births in the population of which the 1000 may be considered as a random sample?

39. It is thought that two physical quantities x and y should be connected by a relation of the form $y = ax^n$. The experimental values are

x	0.5	1.5	2.5	5.0	10.0
y	3.4	7.0	12.8	29.8	68.2

Find the best values of a and n . *Hint.* Fit a straight line to the values of $\log y$ and $\log x$. If Y is the theoretical value of y , $\log Y = \log a + n \log x$.

40. In a survey made in Iowa, random samples of housewives, divided into rural and urban, were asked whether they had done any canning of fruit and vegetables during the previous season. The results are shown in the following table:

	Rural	Urban
Done	357	274
Not done	13	26

Does this indicate a significant difference between rural and urban housewives in respect of their canning operations?

Ans. Yes.

41. A point X is taken at random in a straight line segment AB whose middle point is O . What is the probability that AX , BX , and AO can form a triangle? What fundamental assumption is made in the solution?

Hint. Any two sides of a triangle are together greater than the third side. If the line AB is of length $2a$ and if $AX = x$, then for $x < a$ the condition is that $a + x > 2a - x$. If $x > a$, the condition is that $2a - x + a > x$. The conditions are satisfied if the point x lies between the mid-points of AO and OB .

42. Suppose it costs one cent to draw each individual of a sample. It is required to draw a sample of N from an infinite population in which $\sigma/\mu = 0.1$ and N is to be so large that the probability that the sample mean differs from the population mean by more than 0.1 per cent of the latter is less than 0.01. How much will this sample cost? How much extra will it cost to double the accuracy (that is, to replace 0.1 per cent by 0.05 per cent)?

Ans. \$670, \$2010.

43. Bacterial counts on 15 plates, made with the same dilution of a culture, were as follows: 193, 168, 161, 153, 152, 171, 156, 159, 140, 183, 151, 152, 133, 164, 157. Is the variability consistent with what would be expected if the numbers follow a Poisson law?

Hint. For a Poisson law the variance is equal to the mean. The sampling variance of \bar{x} in a sample of N is such that $N s^2/\sigma^2$ is distributed like χ^2 with $N - 1$ degrees of freedom. Use the mean of the sample as an estimate of σ^2 and compute χ^2 from the sample variance.

44. Show that the least squares condition, $\sum (Z - z)^2 = \min.$, gives, when $Z = a + bx + cy$, the normal equations (16.37).

Hint. Calculus students will differentiate $\sum (Z - z)^2$ partially with respect to a , b , and c . Others can use the method of §14.9, expressing $\sum (Z - z)^2$ as a quadratic in either a , b , or c .

APPENDIX

TABLE I. ORDINATES AND AREAS OF THE NORMAL CURVE, $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

z	$\phi(z)$	$\int_0^z \phi(z) dz$	z	$\phi(z)$	$\int_0^z \phi(z) dz$	z	$\phi(z)$	$\int_0^z \phi(z) dz$
.00	.39894	.00000	.45	.36053	.17364	.90	.26609	.31594
.01	.39892	.00399	.46	.35889	.17724	.91	.26369	.31859
.02	.39886	.00798	.47	.35723	.18082	.92	.26129	.32121
.03	.39876	.01197	.48	.35553	.18439	.93	.25888	.32381
.04	.39862	.01595	.49	.35381	.18793	.94	.25647	.32639
.05	.39844	.01994	.50	.35207	.19146	.95	.25406	.32894
.06	.39822	.02392	.51	.35029	.19497	.96	.25164	.33147
.07	.39797	.02790	.52	.34849	.19847	.97	.24923	.33398
.08	.39767	.03188	.53	.34667	.20194	.98	.24681	.33646
.09	.39733	.03586	.54	.34482	.20540	.99	.24439	.33891
.10	.39695	.03983	.55	.34294	.20884	1.00	.24197	.34134
.11	.39654	.04380	.56	.34105	.21226	1.01	.23955	.34375
.12	.39608	.04776	.57	.33912	.21566	1.02	.23713	.34614
.13	.39559	.05172	.58	.33718	.21904	1.03	.23471	.34850
.14	.39505	.05567	.59	.33521	.22240	1.04	.23230	.35083
.15	.39448	.05962	.60	.33322	.22575	1.05	.22988	.35314
.16	.39387	.06356	.61	.33121	.22907	1.06	.22747	.35543
.17	.39322	.06749	.62	.32918	.23237	1.07	.22506	.35769
.18	.39253	.07142	.63	.32713	.23565	1.08	.22265	.35993
.19	.39181	.07535	.64	.32506	.23891	1.09	.22025	.36214
.20	.39104	.07926	.65	.32297	.24215	1.10	.21785	.36433
.21	.39024	.08317	.66	.32086	.24537	1.11	.21546	.36650
.22	.38940	.08706	.67	.31874	.24857	1.12	.21307	.36864
.23	.38853	.09095	.68	.31659	.25175	1.13	.21069	.37076
.24	.38762	.09483	.69	.31443	.25490	1.14	.20831	.37286
.25	.38667	.09871	.70	.31225	.25804	1.15	.20594	.37493
.26	.38568	.10257	.71	.31006	.26115	1.16	.20357	.37698
.27	.38466	.10642	.72	.30785	.26424	1.17	.20121	.37900
.28	.38361	.11026	.73	.30563	.26730	1.18	.19886	.38100
.29	.38251	.11409	.74	.30339	.27035	1.19	.19652	.38298
.30	.38139	.11791	.75	.30114	.27337	1.20	.19419	.38493
.31	.38023	.12172	.76	.29887	.27637	1.21	.19186	.38686
.32	.37903	.12552	.77	.29659	.27935	1.22	.18954	.38877
.33	.37780	.12930	.78	.29431	.28230	1.23	.18724	.39065
.34	.37654	.13307	.79	.29200	.28524	1.24	.18494	.39251
.35	.37524	.13683	.80	.28969	.28814	1.25	.18265	.39435
.36	.37391	.14058	.81	.28737	.29103	1.26	.18037	.39617
.37	.37255	.14431	.82	.28504	.29389	1.27	.17810	.39796
.38	.37115	.14803	.83	.28269	.29673	1.28	.17585	.39973
.39	.36973	.15173	.84	.28034	.29955	1.29	.17360	.40147
.40	.36827	.15542	.85	.27798	.30234	1.30	.17137	.40320
.41	.36678	.15910	.86	.27562	.30511	1.31	.16915	.40490
.42	.36526	.16276	.87	.27324	.30785	1.32	.16694	.40658
.43	.36371	.16640	.88	.27086	.31057	1.33	.16474	.40824
.44	.36213	.17003	.89	.26848	.31327	1.34	.16256	.40988

TABLE I. ORDINATES AND AREAS OF THE NORMAL CURVE, $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

z	$\phi(z)$	$\int_0^z \phi(z) dz$	z	$\phi(z)$	$\int_0^z \phi(z) dz$	z	$\phi(z)$	$\int_0^z \phi(z) dz$
1.35	.16038	.41149	1.80	.07895	.46407	2.25	.03174	.48778
1.36	.15822	.41309	1.81	.07754	.46485	2.26	.03103	.48809
1.37	.15608	.41466	1.82	.07614	.46562	2.27	.03034	.48840
1.38	.15395	.41621	1.83	.07477	.46638	2.28	.02965	.48870
1.39	.15183	.41774	1.84	.07341	.46712	2.29	.02898	.48899
1.40	.14973	.41924	1.85	.07206	.46784	2.30	.02833	.48928
1.41	.14764	.42073	1.86	.07074	.46856	2.31	.02768	.48956
1.42	.14556	.42220	1.87	.06943	.46926	2.32	.02705	.48983
1.43	.14350	.42364	1.88	.06814	.46995	2.33	.02643	.49010
1.44	.14146	.42507	1.89	.06687	.47062	2.34	.02582	.49036
1.45	.13943	.42647	1.90	.06562	.47128	2.35	.02522	.49061
1.46	.13742	.42786	1.91	.06439	.47193	2.36	.02463	.49086
1.47	.13542	.42922	1.92	.06316	.47257	2.37	.02406	.49111
1.48	.13344	.43056	1.93	.06195	.47320	2.38	.02349	.49134
1.49	.13147	.43189	1.94	.06077	.47381	2.39	.02294	.49158
1.50	.12952	.43319	1.95	.05959	.47441	2.40	.02239	.49180
1.51	.12758	.43448	1.96	.05844	.47500	2.41	.02186	.49202
1.52	.12566	.43574	1.97	.05730	.47558	2.42	.02134	.49224
1.53	.12376	.43699	1.98	.05618	.47615	2.43	.02083	.49245
1.54	.12188	.43822	1.99	.05508	.47670	2.44	.02033	.49266
1.55	.12001	.43943	2.00	.05399	.47725	2.45	.01984	.49286
1.56	.11816	.44062	2.01	.05292	.47778	2.46	.01936	.49305
1.57	.11632	.44179	2.02	.05186	.47831	2.47	.01889	.49324
1.58	.11450	.44295	2.03	.05082	.47882	2.48	.01842	.49343
1.59	.11270	.44408	2.04	.04980	.47932	2.49	.01797	.49361
1.60	.11092	.44520	2.05	.04879	.47982	2.50	.01753	.49379
1.61	.10915	.44630	2.06	.04780	.48030	2.51	.01709	.49396
1.62	.10741	.44738	2.07	.04682	.48077	2.52	.01667	.49413
1.63	.10567	.44845	2.08	.04586	.48124	2.53	.01625	.49430
1.64	.10396	.44950	2.09	.04491	.48169	2.54	.01585	.49446
1.65	.10226	.45053	2.10	.04398	.48214	2.55	.01545	.49461
1.66	.10059	.45154	2.11	.04307	.48257	2.56	.01506	.49477
1.67	.09893	.45254	2.12	.04217	.48300	2.57	.01468	.49492
1.68	.09728	.45352	2.13	.04128	.48341	2.58	.01431	.49506
1.69	.09566	.45449	2.14	.04041	.48382	2.59	.01394	.49520
1.70	.09405	.45543	2.15	.03955	.48422	2.60	.01358	.49534
1.71	.09246	.45637	2.16	.03871	.48461	2.61	.01323	.49547
1.72	.09089	.45728	2.17	.03788	.48500	2.62	.01289	.49560
1.73	.08933	.45818	2.18	.03706	.48537	2.63	.01256	.49573
1.74	.08780	.45907	2.19	.03626	.48574	2.64	.01223	.49585
1.75	.08628	.45994	2.20	.03547	.48610	2.65	.01191	.49598
1.76	.08478	.46080	2.21	.03470	.48645	2.66	.01160	.49609
1.77	.08329	.46164	2.22	.03394	.48679	2.67	.01130	.49621
1.78	.08183	.46246	2.23	.03319	.48713	2.68	.01100	.49632
1.79	.08038	.46327	2.24	.03246	.48745	2.69	.01071	.49643

TABLE I. ORDINATES AND AREAS OF THE NORMAL CURVE, $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

z	$\phi(z)$	$\int_0^z \phi(z) dz$	z	$\phi(z)$	$\int_0^z \phi(z) dz$	z	$\phi(z)$	$\int_0^z \phi(z) dz$
2.70	.01042	.49653	3.15	.00279	.49918	3.60	.00061	.49984
2.71	.01014	.49664	3.16	.00271	.49921	3.61	.00059	.49985
2.72	.00987	.49674	3.17	.00262	.49924	3.62	.00057	.49985
2.73	.00961	.49683	3.18	.00254	.49926	3.63	.00055	.49986
2.74	.00935	.49693	3.19	.00246	.49929	3.64	.00053	.49986
2.75	.00909	.49702	3.20	.00238	.49931	3.65	.00051	.49987
2.76	.00885	.49711	3.21	.00231	.49934	3.66	.00049	.49987
2.77	.00861	.49720	3.22	.00224	.49936	3.67	.00047	.49988
2.78	.00837	.49728	3.23	.00216	.49938	3.68	.00046	.49988
2.79	.00814	.49736	3.24	.00210	.49940	3.69	.00044	.49989
2.80	.00792	.49744	3.25	.00203	.49942	3.70	.00042	.49989
2.81	.00770	.49752	3.26	.00196	.49944	3.71	.00041	.49990
2.82	.00748	.49760	3.27	.00190	.49946	3.72	.00039	.49990
2.83	.00727	.49767	3.28	.00184	.49948	3.73	.00038	.49990
2.84	.00707	.49774	3.29	.00178	.49950	3.74	.00037	.49991
2.85	.00687	.49781	3.30	.00172	.49952	3.75	.00035	.49991
2.86	.00668	.49788	3.31	.00167	.49953	3.76	.00034	.49992
2.87	.00649	.49795	3.32	.00161	.49955	3.77	.00033	.49992
2.88	.00631	.49801	3.33	.00156	.49957	3.78	.00031	.49992
2.89	.00613	.49807	3.34	.00151	.49958	3.79	.00030	.49992
2.90	.00595	.49813	3.35	.00146	.49960	3.80	.00029	.49993
2.91	.00578	.49819	3.36	.00141	.49961	3.81	.00028	.49993
2.92	.00562	.49825	3.37	.00136	.49962	3.82	.00027	.49993
2.93	.00545	.49831	3.38	.00132	.49964	3.83	.00026	.49994
2.94	.00530	.49836	3.39	.00127	.49965	3.84	.00025	.49994
2.95	.00514	.49841	3.40	.00123	.49966	3.85	.00024	.49994
2.96	.00499	.49846	3.41	.00119	.49968	3.86	.00023	.49994
2.97	.00485	.49851	3.42	.00115	.49969	3.87	.00022	.49995
2.98	.00471	.49856	3.43	.00111	.49970	3.88	.00021	.49995
2.99	.00457	.49861	3.44	.00107	.49971	3.89	.00021	.49995
3.00	.00443	.49865	3.45	.00104	.49972	3.90	.00020	.49995
3.01	.00430	.49869	3.46	.00100	.49973	3.91	.00019	.49995
3.02	.00417	.49874	3.47	.00097	.49974	3.92	.00018	.49996
3.03	.00405	.49878	3.48	.00094	.49975	3.93	.00018	.49996
3.04	.00393	.49882	3.49	.00090	.49976	3.94	.00017	.49996
3.05	.00381	.49886	3.50	.00087	.49977	3.95	.00016	.49996
3.06	.00370	.49889	3.51	.00084	.49978	3.96	.00016	.49996
3.07	.00358	.49893	3.52	.00081	.49978	3.97	.00015	.49996
3.08	.00348	.49897	3.53	.00079	.49979	3.98	.00014	.49997
3.09	.00337	.49900	3.54	.00076	.49980	3.99	.00014	.49997
3.10	.00327	.49903	3.55	.00073	.49981			
3.11	.00317	.49906	3.56	.00071	.49981			
3.12	.00307	.49910	3.57	.00068	.49982			
3.13	.00298	.49913	3.58	.00066	.49983			
3.14	.00288	.49916	3.59	.00063	.49983			



TABLE II. VALUES OF t CORRESPONDING TO GIVEN PROBABILITIES *

Degrees of freedom n	Probability of a deviation greater than t					
	.005	.01	.025	.05	.1	.15
1	63.657	31.821	12.706	6.314	3.078	1.963
2	9.925	6.965	4.303	2.920	1.886	1.386
3	5.841	4.541	3.182	2.353	1.638	1.250
4	4.604	3.747	2.776	2.132	1.533	1.190
5	4.032	3.365	2.571	2.015	1.476	1.156
6	3.707	3.143	2.447	1.943	1.440	1.134
7	3.499	2.998	2.365	1.895	1.415	1.119
8	3.355	2.896	2.306	1.860	1.397	1.108
9	3.250	2.821	2.262	1.833	1.383	1.100
10	3.169	2.764	2.228	1.812	1.372	1.093
11	3.106	2.718	2.201	1.796	1.363	1.088
12	3.055	2.681	2.179	1.782	1.356	1.083
13	3.012	2.650	2.160	1.771	1.350	1.079
14	2.977	2.624	2.145	1.761	1.345	1.076
15	2.947	2.602	2.131	1.753	1.341	1.074
16	2.921	2.583	2.120	1.746	1.337	1.071
17	2.898	2.567	2.110	1.740	1.333	1.069
18	2.878	2.552	2.101	1.734	1.330	1.067
19	2.861	2.539	2.093	1.729	1.328	1.066
20	2.845	2.528	2.086	1.725	1.325	1.064
21	2.831	2.518	2.080	1.721	1.323	1.063
22	2.819	2.508	2.074	1.717	1.321	1.061
23	2.807	2.500	2.069	1.714	1.319	1.060
24	2.797	2.492	2.064	1.711	1.318	1.059
25	2.787	2.485	2.060	1.708	1.316	1.058
26	2.779	2.479	2.056	1.706	1.315	1.058
27	2.771	2.473	2.052	1.703	1.314	1.057
28	2.763	2.467	2.048	1.701	1.313	1.056
29	2.756	2.462	2.045	1.699	1.311	1.055
30	2.750	2.457	2.042	1.697	1.310	1.055
∞	2.576	2.326	1.960	1.645	1.282	1.036

The probability of a deviation *numerically* greater than t is twice the probability given at the head of the table.

* This table is reproduced from "Statistical Methods for Research Workers," with the generous permission of the author, Professor R. A. Fisher, and the publishers, Messrs. Oliver and Boyd.

TABLE II. VALUES OF t CORRESPONDING TO GIVEN PROBABILITIES (*cont.*)

Degrees of freedom n	Probability of a deviation greater than t					
	.2	.25	.3	.35	.4	.45
1	1.376	1.000	.727	.510	.325	.158
2	1.061	.816	.617	.445	.289	.142
3	.978	.765	.584	.424	.277	.137
4	.941	.741	.569	.414	.271	.134
5	.920	.727	.559	.408	.267	.132
6	.906	.718	.553	.404	.265	.131
7	.896	.711	.549	.402	.263	.130
8	.889	.706	.546	.399	.262	.130
9	.883	.703	.543	.398	.261	.129
10	.879	.700	.542	.397	.260	.129
11	.876	.697	.540	.396	.260	.129
12	.873	.695	.539	.395	.259	.128
13	.870	.694	.538	.394	.259	.128
14	.868	.692	.537	.393	.258	.128
15	.866	.691	.536	.393	.258	.128
16	.865	.690	.535	.392	.258	.128
17	.863	.689	.534	.392	.257	.128
18	.862	.688	.534	.392	.257	.127
19	.861	.688	.533	.391	.257	.127
20	.860	.687	.533	.391	.257	.127
21	.859	.686	.532	.391	.257	.127
22	.858	.686	.532	.390	.256	.127
23	.858	.685	.532	.390	.256	.127
24	.857	.685	.531	.390	.256	.127
25	.856	.684	.531	.390	.256	.127
26	.856	.684	.531	.390	.256	.127
27	.855	.684	.531	.389	.256	.127
28	.855	.683	.530	.389	.256	.127
29	.854	.683	.530	.389	.256	.127
30	.854	.683	.530	.389	.256	.127
∞	.842	.674	.524	.385	.253	.126

The probability of a deviation *numerically* greater than t is twice the probability given at the head of the table.

TABLE III. VALUES OF χ^2 CORRESPONDING TO GIVEN PROBABILITIES *

Degrees of freedom n	Probability of a deviation greater than χ^2						
	.01	.02	.05	.10	.20	.30	.50
1	6.635	5.412	3.841	2.706	1.642	1.074	.455
2	9.210	7.824	5.991	4.605	3.219	2.408	1.386
3	11.341	9.837	7.815	6.251	4.642	3.665	2.366
4	13.277	11.668	9.488	7.779	5.989	4.878	3.357
5	15.086	13.388	11.070	9.236	7.289	6.064	4.351
6	16.812	15.033	12.592	10.645	8.558	7.231	5.348
7	18.475	16.622	14.067	12.017	9.803	8.383	6.346
8	20.090	18.168	15.507	13.362	11.030	9.524	7.344
9	21.666	19.679	16.919	14.684	12.242	10.656	8.343
10	23.209	21.161	18.307	15.987	13.442	11.781	9.342
11	24.725	22.618	19.675	17.275	14.631	12.899	10.341
12	26.217	24.054	21.026	18.549	15.812	14.011	11.340
13	27.688	25.472	22.362	19.812	16.985	15.119	12.340
14	29.141	26.873	23.685	21.064	18.151	16.222	13.339
15	30.578	28.259	24.996	22.307	19.311	17.322	14.339
16	32.000	29.633	26.296	23.542	20.465	18.418	15.338
17	33.409	30.995	27.587	24.769	21.615	19.511	16.338
18	34.805	32.346	28.869	25.989	22.760	20.601	17.338
19	36.191	33.687	30.144	27.204	23.900	21.689	18.338
20	37.566	35.020	31.410	28.412	25.038	22.775	19.337
21	38.932	36.343	32.671	29.615	26.171	23.858	20.337
22	40.289	37.659	33.924	30.813	27.301	24.939	21.337
23	41.638	38.968	35.172	32.007	28.429	26.018	22.337
24	42.980	40.270	36.415	33.196	29.553	27.096	23.337
25	44.314	41.566	37.652	34.382	30.675	28.172	24.337
26	45.642	42.856	38.885	35.563	31.795	29.246	25.336
27	46.963	44.140	40.113	36.741	32.912	30.319	26.336
28	48.278	45.419	41.337	37.916	34.027	31.391	27.336
29	49.588	46.693	42.557	39.087	35.139	32.461	28.336
30	50.892	47.962	43.773	40.256	36.250	33.530	29.336

For larger values of n , the quantity $(2\chi^2)^{1/2} - (2n - 1)^{1/2}$ may be used as a normal deviate with unit standard deviation.

* This table is reproduced from "Statistical Methods for Research Workers," with the generous permission of the author, Professor R. A. Fisher, and the publishers, Messrs. Oliver and Boyd.

TABLE III. VALUES OF χ^2 CORRESPONDING TO GIVEN PROBABILITIES (cont.)

Degrees of freedom <i>n</i>	Probability of a deviation greater than χ^2					
	.70	.80	.90	.95	.98	.99
1	.148	.0642	.0158	.00393	.000628	.000157
2	.713	.446	.211	.103	.0404	.0201
3	1.424	1.005	.584	.352	.185	.115
4	2.195	1.649	1.064	.711	.429	.297
5	3.000	2.343	1.610	1.145	.752	.554
6	3.828	3.070	2.204	1.635	1.134	.872
7	4.671	3.822	2.833	2.167	1.564	1.239
8	5.527	4.594	3.490	2.733	2.032	1.646
9	6.393	5.380	4.168	3.325	2.532	2.088
10	7.267	6.179	4.865	3.940	3.059	2.558
11	8.148	6.989	5.578	4.575	3.609	3.053
12	9.034	7.807	6.304	5.226	4.178	3.571
13	9.926	8.634	7.042	5.892	4.765	4.107
14	10.821	9.467	7.790	6.571	5.368	4.660
15	11.721	10.307	8.547	7.261	5.985	5.229
16	12.624	11.152	9.312	7.962	6.614	5.812
17	13.531	12.002	10.085	8.672	7.255	6.408
18	14.440	12.857	10.865	9.390	7.906	7.015
19	15.352	13.716	11.651	10.117	8.567	7.633
20	16.266	14.578	12.443	10.851	9.237	8.260
21	17.182	15.445	13.240	11.591	9.915	8.897
22	18.101	16.314	14.041	12.338	10.600	9.542
23	19.021	17.187	14.848	13.091	11.293	10.196
24	19.943	18.062	15.659	13.848	11.992	10.856
25	20.867	18.940	16.473	14.611	12.697	11.524
26	21.792	19.820	17.292	15.379	13.409	12.198
27	22.719	20.703	18.114	16.151	14.125	12.879
28	23.647	21.588	18.939	16.928	14.847	13.565
29	24.577	22.475	19.768	17.708	15.574	14.256
30	25.508	23.364	20.599	18.493	16.306	14.953

For larger values of n , the quantity $(2\chi^2)^{1/2} - (2n - 1)^{1/2}$ may be used as a normal deviate with unit standard deviation.

TABLE IV.* 5% (ROMAN TYPE) AND 1% (BOLD FACE TYPE) POINTS FOR THE DISTRIBUTION OF F

		α, degrees of freedom (for greater mean square)																									
		1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞		
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254	254		
	4,052	4,999	5,493	5,625	5,764	5,859	5,928	5,981	6,022	6,056	6,082	6,106	6,124	6,142	6,169	6,208	6,234	6,258	6,286	6,302	6,323	6,334	6,352	6,361	6,366		
2	18	51	19	00	19	25	19	30	19	33	19	36	19	37	19	38	19	39	19	40	19	41	19	40	19		
	98.49	99.01	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.41	99.42	99.43	99.44	99.45	99.46	99.47	99.48	99.48	99.49	99.49	99.49	99.50	99.50	99.50		
3	10	13	9	55	9	28	9	12	8	84	8	81	8	78	8	76	8	75	8	74	8	73	8	72	8		
	34.12	30.81	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.92	26.83	26.69	26.60	26.50	26.41	26.35	26.27	26.23	26.18	26.14	26.12	26.12		
4	7	7	1	6	94	6	39	6	26	6	16	6	09	6	04	6	00	5	96	5	93	5	91	5	87		
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37	14.24	14.15	14.02	13.93	13.83	13.74	13.69	13.61	13.57	13.52	13.48	13.46	13.46		
5	6	6	1	5	79	5	41	5	19	5	05	4	95	4	88	4	82	4	78	4	74	4	70	4	68		
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.96	9.89	9.77	9.68	9.55	9.47	9.38	9.29	9.24	9.17	9.13	9.07	9.04	9.02	9.02		
6	5	5	9	4	76	4	57	4	39	4	28	4	21	4	15	4	10	4	06	4	03	4	00	3	96		
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.60	7.52	7.39	7.31	7.23	7.14	7.09	7.02	6.99	6.94	6.90	6.88	6.88		
7	5	5	9	4	74	4	35	4	12	3	97	3	87	3	79	3	73	3	68	3	63	3	60	3	57		
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47	6.35	6.27	6.15	6.07	5.98	5.90	5.85	5.78	5.75	5.70	5.67	5.65	5.65		
8	5	5	3	4	6	3	69	3	58	3	50	3	44	3	39	3	34	3	31	3	28	3	25	3	23		
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67	5.56	5.48	5.36	5.28	5.20	5.11	5.06	5.00	4.96	4.91	4.88	4.86	4.86		
9	5	5	1	2	6	3	63	3	48	3	37	3	29	3	23	3	18	3	13	3	10	3	07	3	02		
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11	5.00	4.92	4.80	4.73	4.64	4.56	4.51	4.45	4.41	4.36	4.33	4.31	4.31		
10	4	4	9	4	10	3	71	3	48	3	33	3	23	3	14	3	07	3	02	2	97	2	94	2	91		
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71	4.60	4.52	4.41	4.33	4.25	4.17	4.12	4.05	4.01	3.96	3.93	3.91	3.91		
11	4	4	4	3	98	3	59	3	36	3	20	3	09	3	01	2	95	2	90	2	86	2	82	2	79		
	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40	4.29	4.21	4.10	4.02	3.94	3.86	3.80	3.74	3.70	3.66	3.62	3.60	3.60		
12	4	4	5	3	88	3	49	3	26	3	11	3	00	2	92	2	85	2	80	2	76	2	72	2	69		
	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16	4.05	3.98	3.86	3.78	3.70	3.61	3.56	3.49	3.46	3.41	3.38	3.36	3.36		
13	4	4	6	3	80	3	41	3	18	3	02	2	92	2	84	2	77	2	72	2	68	2	64	2	61		
	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.85	3.78	3.67	3.59	3.51	3.43	3.37	3.30	3.27	3.21	3.18	3.16	3.16		

* Reproduced from *Statistical Methods* by G. W. Snedecor by permission of the author and the publisher, Collegiate Press, Inc., Ames, Iowa.

TABLE IV. 5% (ROMAN TYPE) AND 1% (BOLD FACE TYPE) POINTS FOR THE DISTRIBUTION OF F

F ₁	n: degrees of freedom (for greater mean square)																						F ₂	
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	100	200	500		∞
14	4.80	3.74	3.24	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
15	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00
16	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
17	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.56	3.48	3.36	3.29	3.20	3.12	3.07	3.00	2.97	2.92	2.89	2.87
18	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.60	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
19	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55	3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.89	2.86	2.83	2.77	2.75
20	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
21	8.40	6.11	5.16	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45	3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65
22	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
23	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37	3.27	3.19	3.07	3.00	2.91	2.83	2.78	2.71	2.68	2.62	2.59	2.57
24	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	1.99	1.96	1.94	1.91	1.90	1.88
25	8.16	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.19	3.12	3.00	2.92	2.84	2.76	2.70	2.63	2.60	2.54	2.51	2.49
26	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
27	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23	3.13	3.05	2.94	2.85	2.77	2.69	2.63	2.56	2.53	2.47	2.44	2.42
28	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
29	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24	3.17	3.07	2.99	2.88	2.80	2.72	2.63	2.58	2.51	2.47	2.42	2.38	2.36
30	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
31	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.37	2.33	2.31
32	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
33	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.97	2.89	2.78	2.70	2.62	2.53	2.48	2.41	2.37	2.32	2.28	2.26
34	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
35	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.21
36	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
37	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05	2.99	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19	2.17
38	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69
39	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96	2.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15	2.13

TABLE IV 5% (ROMAN TYPE) AND 1% (BOLD FACE TYPE) POINTS FOR THE DISTRIBUTION OF F

m	n: degrees of freedom (for greater mean square)																								m
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67	27
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93	2.83	2.74	2.63	2.55	2.47	2.38	2.33	2.25	2.21	2.16	2.12	2.10	
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65	28
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90	2.80	2.71	2.60	2.52	2.44	2.35	2.30	2.22	2.18	2.13	2.09	2.06	
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64	29
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87	2.77	2.68	2.57	2.49	2.41	2.32	2.27	2.19	2.15	2.10	2.06	2.03	
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62	30
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84	2.74	2.66	2.55	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03	2.01	
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59	32
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80	2.70	2.62	2.51	2.42	2.34	2.25	2.20	2.12	2.08	2.02	1.98	1.96	
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57	34
	7.44	5.29	4.42	3.93	3.61	3.38	3.21	3.08	2.97	2.89	2.82	2.76	2.66	2.58	2.47	2.38	2.30	2.21	2.15	2.08	2.04	1.98	1.94	1.91	
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55	36
	7.39	5.25	4.38	3.89	3.58	3.35	3.18	3.04	2.94	2.86	2.78	2.72	2.62	2.54	2.43	2.35	2.26	2.17	2.12	2.04	2.00	1.94	1.90	1.87	
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53	38
	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.91	2.82	2.75	2.69	2.59	2.51	2.40	2.32	2.22	2.14	2.08	2.00	1.97	1.90	1.86	1.84	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51	40
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66	2.56	2.49	2.37	2.29	2.20	2.11	2.05	1.97	1.94	1.88	1.84	1.81	
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49	42
	7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70	2.64	2.54	2.46	2.35	2.26	2.17	2.08	2.02	1.94	1.91	1.85	1.80	1.78	
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48	44
	7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68	2.62	2.52	2.44	2.32	2.24	2.15	2.06	2.00	1.92	1.88	1.82	1.78	1.75	
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46	46
	7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66	2.60	2.50	2.42	2.30	2.22	2.13	2.04	1.98	1.91	1.86	1.80	1.76	1.73	
48	4.04	3.19	2.80	2.56	2.41	2.29	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45	48
	7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64	2.58	2.48	2.40	2.28	2.20	2.11	2.02	1.96	1.88	1.84	1.78	1.73	1.70	

TABLE IV. 5% (ROMAN TYPE) AND 1% (BOLD FACE TYPE) POINTS FOR THE DISTRIBUTION OF F

n ₂	n ₁ degrees of freedom (for greater mean square)																							∞
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
55	7.17	5.06	4.26	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56	2.46	2.39	2.26	2.18	2.10	2.00	1.94	1.86	1.82	1.76	1.71	1.68
	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
60	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.59	2.53	2.43	2.35	2.23	2.15	2.06	1.96	1.90	1.82	1.78	1.71	1.66	1.64
	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
65	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.40	2.32	2.20	2.12	2.03	1.93	1.87	1.79	1.74	1.68	1.63	1.60
	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
70	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.79	2.70	2.61	2.54	2.47	2.37	2.30	2.18	2.09	2.00	1.90	1.84	1.76	1.71	1.64	1.60	1.56
	2.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
80	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45	2.35	2.28	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.62	1.56	1.53
	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
100	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.56	2.48	2.41	2.32	2.24	2.11	2.03	1.94	1.84	1.78	1.70	1.65	1.57	1.52	1.49
	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.43	1.39	1.34	1.30	1.28
125	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36	2.26	2.19	2.06	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43
	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
150	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.65	2.56	2.47	2.40	2.33	2.23	2.15	2.03	1.94	1.85	1.75	1.68	1.59	1.54	1.46	1.40	1.37
	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.23
200	6.81	4.75	3.91	3.44	3.14	2.92	2.76	2.62	2.53	2.44	2.37	2.30	2.20	2.12	2.00	1.91	1.83	1.72	1.66	1.56	1.51	1.43	1.37	1.33
	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
400	6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.34	2.28	2.17	2.09	1.97	1.88	1.79	1.69	1.62	1.53	1.48	1.39	1.33	1.28
	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
1000	6.70	4.66	3.83	3.36	3.06	2.85	2.69	2.55	2.46	2.37	2.29	2.23	2.12	2.04	1.92	1.84	1.74	1.64	1.57	1.47	1.42	1.32	1.24	1.19
	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
∞	6.66	4.62	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26	2.20	2.09	2.01	1.89	1.81	1.71	1.61	1.54	1.44	1.38	1.28	1.19	1.11
	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.06
	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18	2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.08

TABLE V. RANDOM SAMPLING NUMBERS*

First Thousand										
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	23 15	75 48	59 01	83 72	59 93	76 24	97 08	86 95	23 03	67 44
2	05 54	55 50	43 10	53 74	35 08	90 61	18 37	44 10	96 22	13 43
3	14 87	16 03	50 32	40 43	62 23	50 05	10 03	22 11	54 38	08 34
4	38 97	67 49	51 94	05 17	58 53	78 80	59 01	94 32	42 87	16 95
5	97 31	26 17	18 99	75 53	08 70	94 25	12 58	41 54	88 21	05 13
6	11 74	26 93	81 44	33 93	08 72	32 79	73 31	18 22	64 70	68 50
7	43 36	12 88	59 11	01 64	56 23	93 00	90 04	99 43	64 07	40 36
8	93 80	62 04	78 38	26 80	44 91	55 75	11 89	32 58	47 55	25 71
9	49 54	01 31	81 08	42 98	41 87	69 53	82 96	61 77	73 80	95 27
10	36 76	87 26	33 37	94 82	15 69	41 95	96 86	70 45	27 48	38 80
11	07 09	25 23	92 24	62 71	26 07	06 55	84 53	44 67	33 84	53 20
12	43 31	00 10	81 44	86 38	03 07	52 55	51 61	48 89	74 29	46 47
13	61 57	00 63	60 06	17 36	37 75	63 14	89 51	23 35	01 74	69 93
14	31 35	28 37	99 10	77 91	89 41	31 57	97 64	48 62	58 48	69 19
15	57 04	88 65	26 27	79 59	36 82	90 52	95 65	46 35	06 53	22 54
16	09 24	34 42	00 68	72 10	71 37	30 72	97 57	56 09	29 82	76 50
17	97 95	53 50	18 40	89 48	83 29	52 23	08 25	21 22	53 26	15 87
18	93 73	25 95	70 43	78 19	88 85	56 67	16 68	26 95	99 64	45 69
19	72 62	11 12	25 00	92 26	82 64	35 66	65 94	34 71	68 75	18 67
20	61 02	07 44	18 45	37 12	07 94	95 91	73 78	66 99	53 61	93 78
21	97 83	98 54	74 33	05 59	17 18	45 47	35 41	44 22	03 42	30 00
22	89 16	09 71	92 22	23 29	06 37	35 05	54 54	89 88	43 81	63 61
23	25 96	68 82	20 62	87 17	92 65	02 82	35 28	62 84	91 95	48 83
24	81 44	33 17	19 05	04 95	48 06	74 69	00 75	67 65	01 71	65 45
25	11 32	25 49	31 42	36 23	43 86	08 62	49 76	67 42	24 52	32 45
Second Thousand										
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	64 75	58 38	85 84	12 22	59 20	17 69	61 56	55 95	04 59	59 47
2	10 30	25 22	89 77	43 63	44 30	38 11	24 90	67 07	34 82	33 28
3	71 01	79 84	95 51	30 85	03 74	66 59	10 28	87 53	76 56	91 49
4	60 01	25 56	05 88	41 03	48 79	79 65	59 01	69 78	80 00	36 66
5	37 33	09 46	56 49	16 14	28 02	48 27	45 47	55 44	55 36	50 90
6	47 86	98 70	01 31	59 11	22 73	60 62	61 28	22 34	69 16	12 12
7	38 04	04 27	37 64	16 78	95 78	39 32	34 93	24 88	43 43	87 06
8	73 50	83 09	08 83	05 48	00 78	36 66	93 02	95 56	46 04	53 36
9	32 62	34 64	74 84	06 10	43 24	20 62	83 73	19 32	35 64	39 69
10	97 59	19 95	49 36	63 03	51 06	62 06	99 29	75 95	32 05	77 34
11	74 01	23 19	55 59	79 09	69 82	66 22	42 40	15 96	74 90	75 89
12	56 75	42 64	57 13	35 10	50 14	90 96	63 36	74 69	09 63	34 88
13	49 80	04 99	08 54	83 12	19 98	08 52	82 63	72 92	92 36	50 26
14	43 58	48 96	47 24	87 85	66 70	00 22	15 01	93 99	59 16	23 77
15	16 65	37 96	64 60	32 57	13 01	35 74	28 36	36 73	05 88	72 29
16	48 50	26 90	55 65	32 25	87 48	31 44	68 02	37 31	25 29	63 67
17	96 76	55 46	92 36	31 68	62 30	48 29	63 83	52 23	81 66	40 94
18	38 92	36 15	50 80	35 78	17 84	23 44	41 24	63 33	99 22	81 28
19	77 95	88 16	94 25	22 50	55 87	51 07	30 10	70 60	21 86	19 61
20	17 92	82 80	65 25	58 60	87 71	02 64	18 50	64 65	79 64	81 70
21	94 03	68 59	78 02	31 80	44 99	41 05	41 05	31 87	43 12	15 96
22	47 46	06 04	79 56	23 04	84 17	14 37	28 51	67 27	55 80	03 68
23	47 85	65 60	88 51	99 28	24 39	40 64	41 71	70 13	46 31	82 88
24	57 61	63 46	53 92	29 86	20 18	10 37	57 65	15 62	98 69	07 56
25	08 30	09 27	04 66	75 26	66 10	57 18	87 91	07 54	22 22	20 13

* Reproduced with the permission of Professor E. S. Pearson from M. G. Kendall and B. Babington Smith, *Tables of Random Sampling Numbers* (Tracts for Computers, No. 24), Cambridge Univ. Press.

TABLE V. RANDOM SAMPLING NUMBERS (cont.)

Third Thousand										
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	89 22	10 23	62 65	78 77	47 33	51 27	23 02	13 92	44 13	96 51
2	04 00	59 98	18 63	91 82	90 32	94 01	24 23	63 01	26 11	06 50
3	98 54	63 80	66 50	85 67	50 45	40 64	52 28	41 53	25 44	41 25
4	41 71	98 44	01 59	22 60	13 14	54 58	14 03	98 49	98 86	55 79
5	28 73	37 24	89 00	78 52	58 43	24 61	34 97	97 85	56 78	44 71
6	65 21	38 39	27 77	76 20	30 86	80 74	22 43	95 68	47 68	37 92
7	65 55	31 26	78 90	90 69	04 66	43 67	02 62	17 69	90 03	12 05
8	05 66	86 90	80 73	02 98	57 46	58 33	27 82	31 45	98 69	29 98
9	39 30	29 97	18 49	75 77	95 19	27 38	77 63	73 47	26 29	16 12
10	64 59	23 22	54 45	87 92	94 31	38 32	00 59	81 18	06 78	71 37
11	07 51	34 87	92 47	31 48	36 60	68 90	70 53	36 82	57 99	15 82
12	86 59	36 85	01 56	63 89	98 00	82 83	93 51	48 56	54 10	72 32
13	83 73	52 25	99 97	97 78	12 48	36 83	89 95	60 32	41 06	76 14
14	08 59	52 18	26 54	65 50	82 04	87 99	01 70	33 56	25 80	63 84
15	41 27	32 71	49 44	29 36	94 58	16 82	86 39	62 15	86 43	54 31
16	00 47	37 59	08 56	23 81	22 42	72 63	17 63	14 47	25 20	63 47
17	86 13	15 37	89 81	38 30	78 68	89 13	29 61	82 07	00 98	64 32
18	33 84	97 83	59 04	40 20	35 86	03 17	68 86	63 08	01 82	25 46
19	61 87	04 16	57 07	46 80	86 12	98 08	39 73	49 20	77 54	50 91
20	43 89	86 59	23 25	07 88	61 29	78 49	19 76	53 91	50 08	07 86
21	29 93	93 91	23 04	54 84	59 85	60 95	20 66	41 28	72 64	64 73
22	38 50	58 55	55 14	38 85	50 77	18 65	79 48	87 67	83 17	08 19
23	31 82	43 84	31 67	12 52	55 11	72 04	41 15	62 53	27 98	22 68
24	91 43	00 37	67 13	56 11	55 97	06 75	09 25	52 02	39 13	87 53
25	38 63	56 89	76 25	49 89	75 26	96 45	80 38	05 04	11 66	35 14

Fourth Thousand										
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	02 49	05 41	22 27	94 43	93 64	04 23	07 20	74 11	67 95	40 82
2	11 96	73 64	69 60	62 78	37 01	09 25	33 02	08 01	38 53	74 82
3	48 25	68 34	65 49	69 92	40 79	05 40	33 51	54 39	61 30	31 36
4	27 24	67 30	80 21	48 12	35 36	04 88	18 99	77 49	48 49	30 71
5	32 53	27 72	65 72	43 07	07 22	86 52	91 84	57 92	65 71	00 11
6	66 75	79 89	55 92	37 59	34 31	43 20	45 58	25 45	44 36	92 65
7	11 26	63 45	45 76	50 59	77 46	34 66	82 69	99 26	74 29	75 16
8	17 87	23 91	42 45	56 18	01 46	93 13	74 89	24 64	25 75	92 84
9	62 56	13 03	65 03	40 81	47 54	51 79	80 81	33 61	01 09	77 30
10	62 79	63 07	79 35	49 77	05 01	30 10	50 81	33 00	99 79	19 70
11	75 51	02 17	71 04	33 93	36 60	42 75	76 22	23 87	56 54	84 68
12	87 43	90 16	91 63	51 72	65 90	44 43	70 72	17 98	70 63	90 32
13	97 74	20 26	21 10	74 87	88 03	38 33	76 52	26 92	14 95	90 51
14	98 81	10 60	01 21	57 10	28 75	21 82	88 39	12 85	18 86	16 24
15	51 26	40 18	52 64	60 79	25 53	29 00	42 66	95 78	58 36	29 98
16	40 23	99 33	76 10	41 96	86 10	49 12	00 29	41 80	03 59	93 17
17	26 93	65 91	86 51	66 72	76 45	46 32	94 46	81 94	19 06	66 47
18	88 50	21 17	16 98	29 94	09 74	42 39	46 22	00 69	09 48	16 46
19	63 49	93 80	93 25	59 36	19 95	79 86	78 05	69 01	02 33	83 74
20	36 37	98 12	06 03	31 77	87 10	73 82	83 10	83 60	50 94	40 91
21	93 80	12 23	22 47	47 95	70 17	59 33	43 06	47 43	06 12	66 60
22	29 85	68 71	20 56	31 15	00 53	25 36	58 12	65 22	41 40	24 31
23	97 72	08 79	31 88	26 51	30 50	71 01	71 51	77 06	95 79	29 19
24	85 23	70 91	05 74	60 14	63 77	59 93	81 56	47 34	17 79	27 53
25	75 74	67 52	68 31	72 79	57 73	72 36	48 73	24 36	87 90	68 02

TABLE V. RANDOM SAMPLING NUMBERS (cont.)

Fifth Thousand										
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	29 93	50 69	71 63	17 55	25 79	10 47	88 93	79 61	42 82	13 63
2	15 11	40 71	26 51	89 07	77 87	75 51	01 31	03 42	94 24	81 11
3	03 87	04 32	25 10	58 98	76 29	22 03	99 41	24 38	12 76	50 22
4	79 39	03 91	88 40	75 64	52 69	65 95	92 06	40 14	28 42	29 60
5	30 03	50 69	15 79	19 65	44 28	64 81	95 23	14 48	72 18	15 94
6	29 03	99 98	61 28	75 97	98 02	68 53	13 91	98 38	13 72	43 73
7	78 19	60 81	08 24	10 74	97 77	09 59	94 35	69 84	82 09	49 56
8	15 84	78 54	93 91	44 29	13 51	80 13	07 37	52 21	53 91	09 86
9	36 61	46 22	48 49	19 49	72 09	92 58	79 20	53 41	02 18	00 64
10	40 54	95 48	84 91	46 54	38 62	35 54	14 44	66 88	89 47	41 80
11	40 87	80 89	97 14	28 60	99 82	90 30	87 80	07 51	58 71	66 58
12	10 22	94 92	82 41	17 33	14 68	59 45	51 87	56 08	90 80	66 60
13	15 91	87 67	87 30	62 42	59 28	44 12	42 50	88 31	13 77	16 14
14	13 40	31 87	96 49	90 99	44 04	64 97	94 14	62 18	15 59	83 35
15	66 52	39 45	96 74	90 89	02 71	10 00	99 86	48 17	64 06	89 09
16	91 66	53 64	69 68	34 31	78 70	25 97	50 46	62 21	27 25	06 20
17	67 41	58 75	15 08	20 77	37 29	73 20	15 75	93 96	91 76	96 99
18	76 52	79 69	96 23	72 43	34 48	63 39	23 23	54 60	88 79	06 17
19	19 81	54 77	89 74	34 81	71 47	10 95	43 43	55 81	19 45	44 07
20	25 59	25 35	87 76	38 47	25 75	84 34	76 89	18 05	73 95	72 22
21	55 90	24 55	39 63	64 63	16 09	95 99	98 28	87 40	66 66	66 92
22	02 47	05 83	76 79	79 42	24 82	42 42	39 61	62 47	49 11	72 64
23	18 63	05 32	63 13	31 99	76 19	35 85	91 23	50 14	63 28	86 59
24	89 67	33 82	30 16	06 39	20 07	59 50	33 84	02 76	45 03	33 33
25	62 98	66 73	64 06	59 51	74 27	84 62	31 45	65 82	86 05	73 00
Sixth Thousand										
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	27 50	13 05	46 34	63 85	87 60	35 55	05 67	88 15	47 00	50 92
2	02 31	57 57	62 98	41 09	66 01	69 88	92 83	35 70	76 59	02 58
3	37 43	12 83	66 39	77 33	63 26	53 99	48 65	23 06	94 29	53 04
4	83 56	65 54	19 33	35 42	92 12	37 14	70 75	18 58	98 57	12 52
5	06 81	56 27	49 32	12 42	92 42	05 96	82 94	70 25	45 49	18 16
6	39 15	03 60	15 56	73 16	48 74	50 27	43 42	58 36	73 16	39 90
7	84 45	71 93	10 27	15 83	84 20	57 42	41 28	42 06	15 90	70 47
8	82 47	05 77	06 89	47 13	92 85	60 12	32 89	25 22	42 38	87 37
9	98 04	06 70	24 21	69 02	65 42	55 33	11 95	72 35	73 23	57 26
10	18 33	49 04	14 33	48 50	15 64	58 26	14 91	46 02	72 13	48 62
11	33 92	19 93	38 27	43 40	27 72	79 74	86 57	41 83	58 71	56 99
12	48 66	74 30	44 81	06 80	29 09	50 31	69 61	24 64	28 89	97 79
13	85 85	07 54	21 50	31 80	10 19	56 65	82 52	26 58	55 12	26 34
14	08 27	08 08	35 87	96 57	33 12	01 77	52 76	09 89	71 12	17 69
15	59 61	22 14	26 09	96 75	17 94	51 08	41 91	45 94	80 48	59 92
16	17 45	77 79	31 66	36 54	92 85	65 60	53 98	63 50	11 20	96 63
17	11 26	37 08	07 71	95 95	39 75	92 48	99 78	23 33	19 56	06 67
18	48 08	13 98	16 52	41 15	73 96	32 55	03 12	38 30	88 77	17 03
19	76 27	72 22	99 61	72 15	00 25	21 54	47 79	18 41	58 50	57 66
20	98 89	22 25	79 92	53 55	07 98	66 71	53 29	61 71	56 96	41 78
21	88 69	61 63	01 67	61 88	58 79	35 65	08 45	63 38	69 86	79 47
22	12 58	13 75	80 98	01 35	91 16	18 36	90 54	99 17	68 36	85 06
23	08 86	96 36	14 09	43 85	51 20	65 18	06 40	52 17	48 10	68 97
24	33 81	05 51	32 48	60 12	32 44	08 12	89 00	98 82	79 17	97 22
25	05 15	99 28	87 15	07 08	66 92	53 81	69 42	02 27	65 33	57 69

TABLE V. RANDOM SAMPLING NUMBERS (cont.)

Seventh Thousand										
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	80 30	23 64	67 96	21 33	36 90	03 91	69 33	90 13	34 48	02 19
2	61 29	89 61	32 08	12 62	26 08	42 00	31 73	31 30	30 61	34 11
3	23 33	61 01	02 21	11 81	51 32	36 10	23 74	50 31	90 11	73 52
4	94 21	32 92	93 50	72 67	23 20	74 59	30 30	48 66	75 32	27 97
5	87 61	92 69	01 60	28 79	74 76	86 06	39 29	73 85	03 27	50 57
6	37 56	19 18	03 42	86 03	85 74	44 81	86 45	71 16	13 52	35 56
7	64 86	66 31	55 04	88 40	10 30	84 38	06 13	58 83	62 04	63 52
8	22 69	58 45	49 23	09 81	98 84	05 04	75 99	27 70	72 79	32 19
9	23 22	14 22	64 90	10 26	74 23	53 91	27 73	78 19	92 43	68 10
10	42 38	59 64	72 96	46 57	89 67	22 81	94 56	69 84	18 31	06 39
11	17 18	01 34	10 98	37 48	93 86	88 59	69 53	78 86	37 26	85 48
12	39 45	69 53	94 89	58 97	29 33	29 19	50 94	80 57	31 99	38 91
13	43 18	11 42	56 19	48 44	45 02	84 29	01 78	65 77	76 84	88 85
14	59 44	06 45	68 55	16 65	66 13	38 00	95 76	50 67	67 65	18 83
15	01 50	34 32	38 00	37 57	47 82	66 59	19 50	87 14	35 59	79 47
16	79 14	60 35	47 95	90 71	31 03	85 37	38 70	34 16	64 55	66 49
17	01 56	63 68	80 26	14 97	23 88	59 22	82 39	70 83	48 34	46 48
18	25 76	18 71	29 25	15 51	92 96	01 01	28 18	03 35	11 10	27 84
19	23 52	10 83	45 06	49 85	35 45	84 08	81 13	52 57	21 23	67 02
20	91 64	08 64	25 74	16 10	97 31	10 27	24 48	89 06	42 81	29 10
21	80 86	07 27	26 70	08 65	85 20	31 23	28 99	39 63	32 03	71 91
22	31 71	37 60	95 60	94 95	54 45	27 97	03 67	30 54	86 04	12 41
23	05 83	50 36	09 04	39 15	66 55	80 36	39 71	24 10	62 22	21 53
24	98 70	02 90	30 63	62 59	26 04	97 20	00 91	28 80	40 23	09 91
25	82 79	35 45	64 53	93 24	86 55	48 72	18 57	05 79	20 09	31 46
Eighth Thousand										
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	37 52	49 55	40 65	27 61	08 59	91 23	26 18	95 04	98 20	99 52
2	48 16	69 65	69 02	08 83	08 83	68 37	00 96	13 59	12 16	17 93
3	50 43	06 59	56 53	30 61	40 21	29 06	49 60	90 38	21 43	19 25
4	89 31	62 79	45 73	71 72	77 11	28 80	72 35	75 77	24 72	98 43
5	63 29	90 61	86 39	07 38	38 85	77 06	10 23	30 84	07 95	30 76
6	71 68	93 94	08 72	36 27	85 89	40 59	83 37	93 85	73 97	84 05
7	05 06	96 63	58 24	05 95	56 64	77 53	85 64	15 95	93 91	59 03
8	03 35	58 95	46 44	25 70	31 66	01 05	44 44	62 91	36 31	45 04
9	13 04	57 67	74 77	53 35	93 51	82 83	27 38	63 16	04 48	75 23
10	49 96	43 94	56 04	02 79	55 78	01 44	75 26	85 54	01 81	32 82
11	24 36	24 08	44 77	57 07	54 41	04 56	09 44	30 58	25 45	37 56
12	55 19	97 20	01 11	47 45	79 79	06 72	12 81	86 97	54 09	06 53
13	02 28	54 60	28 35	32 94	36 74	51 63	96 90	04 13	30 43	10 14
14	90 50	13 78	22 20	37 56	97 95	49 95	91 15	52 73	12 93	78 94
15	33 71	32 43	29 58	47 38	39 96	67 51	64 47	49 91	64 58	93 07
16	70 58	28 49	54 32	97 70	27 81	64 69	71 52	02 56	61 37	04 58
17	09 68	96 10	57 78	85 00	89 81	98 30	19 40	76 28	62 99	99 83
18	19 36	60 85	35 04	12 87	83 88	66 54	32 00	30 20	05 30	42 63
19	04 75	44 49	64 26	51 46	80 50	53 91	00 55	67 36	68 66	08 29
20	79 83	32 39	46 77	56 83	42 21	60 03	14 47	07 01	66 85	49 22
21	80 99	42 43	08 58	54 41	98 05	54 39	34 42	97 47	38 35	59 40
22	48 83	64 99	86 94	48 78	79 20	62 23	56 45	92 65	56 36	83 02
23	28 45	35 85	22 20	13 01	73 96	70 05	84 50	68 59	96 58	16 63
24	52 07	63 15	82 30	66 23	14 26	66 61	17 80	41 97	40 27	24 80
25	39 14	52 18	35 87	48 55	48 81	03 11	26 99	03 80	08 86	50 42

TABLE VI. VALUES OF $\tanh z'$

z'	0	1	2	3	4	5	6	7	8	9
.00	.0000	.0010	.0020	.0030	.0040	.0050	.0060	.0070	.0080	.0090
.01	.0100	.0110	.0120	.0130	.0140	.0150	.0160	.0170	.0180	.0190
.02	.0200	.0210	.0220	.0230	.0240	.0250	.0260	.0270	.0280	.0290
.03	.0300	.0310	.0320	.0330	.0340	.0350	.0360	.0370	.0380	.0390
.04	.0400	.0410	.0420	.0430	.0440	.0450	.0460	.0470	.0480	.0490
.05	.0500	.0510	.0520	.0530	.0539	.0549	.0559	.0569	.0579	.0589
.06	.0599	.0609	.0619	.0629	.0639	.0649	.0659	.0669	.0679	.0689
.07	.0699	.0709	.0719	.0729	.0739	.0749	.0759	.0768	.0778	.0788
.08	.0798	.0808	.0818	.0828	.0838	.0848	.0858	.0868	.0878	.0888
.09	.0898	.0907	.0917	.0927	.0937	.0947	.0957	.0967	.0977	.0987
.10	.0997	.1007	.1016	.1026	.1036	.1046	.1056	.1066	.1076	.1086
.11	.1096	.1105	.1115	.1125	.1135	.1145	.1155	.1165	.1175	.1184
.12	.1194	.1204	.1214	.1224	.1234	.1244	.1253	.1263	.1273	.1283
.13	.1293	.1303	.1312	.1322	.1332	.1342	.1352	.1361	.1371	.1381
.14	.1391	.1401	.1411	.1420	.1430	.1440	.1450	.1460	.1469	.1479
.15	.1489	.1499	.1508	.1518	.1528	.1538	.1547	.1557	.1567	.1577
.16	.1586	.1596	.1606	.1616	.1625	.1635	.1645	.1655	.1664	.1674
.17	.1684	.1694	.1703	.1713	.1723	.1732	.1742	.1752	.1761	.1771
.18	.1781	.1790	.1800	.1810	.1820	.1829	.1839	.1849	.1858	.1868
.19	.1877	.1887	.1897	.1906	.1916	.1926	.1935	.1945	.1955	.1964
.20	.1974	.1983	.1993	.2003	.2012	.2022	.2031	.2041	.2051	.2060
.21	.2070	.2079	.2089	.2098	.2108	.2117	.2127	.2137	.2146	.2156
.22	.2165	.2175	.2184	.2194	.2203	.2213	.2222	.2232	.2241	.2251
.23	.2260	.2270	.2279	.2289	.2298	.2308	.2317	.2327	.2336	.2346
.24	.2355	.2364	.2374	.2383	.2393	.2402	.2412	.2421	.2430	.2440
.25	.2449	.2459	.2468	.2477	.2487	.2496	.2506	.2515	.2524	.2534
.26	.2543	.2552	.2562	.2571	.2580	.2590	.2599	.2608	.2618	.2627
.27	.2636	.2646	.2655	.2664	.2673	.2683	.2692	.2701	.2711	.2720
.28	.2729	.2738	.2748	.2757	.2766	.2775	.2784	.2794	.2803	.2812
.29	.2821	.2831	.2840	.2849	.2858	.2867	.2876	.2886	.2895	.2904
.30	.2913	.2922	.2931	.2941	.2950	.2959	.2968	.2977	.2986	.2995
.31	.3004	.3013	.3023	.3032	.3041	.3050	.3059	.3068	.3077	.3086
.32	.3095	.3104	.3113	.3122	.3131	.3140	.3149	.3158	.3167	.3176
.33	.3185	.3194	.3203	.3212	.3221	.3230	.3239	.3248	.3257	.3266
.34	.3275	.3284	.3293	.3302	.3310	.3319	.3328	.3337	.3346	.3355
.35	.3364	.3373	.3381	.3390	.3399	.3408	.3417	.3426	.3435	.3443
.36	.3452	.3461	.3470	.3479	.3487	.3496	.3505	.3514	.3522	.3531
.37	.3540	.3549	.3557	.3566	.3575	.3584	.3592	.3601	.3610	.3618
.38	.3627	.3636	.3644	.3653	.3662	.3670	.3679	.3688	.3696	.3705
.39	.3714	.3722	.3731	.3739	.3748	.3757	.3766	.3774	.3782	.3791
.40	.3799	.3808	.3817	.3825	.3834	.3842	.3851	.3859	.3868	.3876
.41	.3885	.3893	.3902	.3910	.3919	.3927	.3936	.3944	.3952	.3961
.42	.3969	.3978	.3986	.3995	.4003	.4011	.4020	.4028	.4036	.4045
.43	.4053	.4062	.4070	.4078	.4087	.4095	.4103	.4112	.4120	.4128
.44	.4136	.4145	.4153	.4161	.4170	.4178	.4186	.4194	.4203	.4211
.45	.4219	.4227	.4235	.4244	.4252	.4260	.4268	.4276	.4285	.4293
.46	.4301	.4309	.4317	.4325	.4333	.4342	.4350	.4358	.4366	.4374
.47	.4382	.4390	.4398	.4406	.4414	.4422	.4430	.4438	.4446	.4454
.48	.4462	.4470	.4478	.4486	.4494	.4502	.4510	.4518	.4526	.4534
.49	.4542	.4550	.4558	.4566	.4574	.4582	.4590	.4598	.4605	.4613

Reproduced, by permission of the author, from *Numerical Tables* by J. W. Campbell (Department of Mathematics, University of Alberta, Canada).

TABLE VI. VALUES OF $\tanh z'$ (cont.)

z'	0	1	2	3	4	5	6	7	8	9
.50	.4621	4629	4637	4645	4653	4660	4668	4676	4684	4692
.51	.4699	4707	4715	4723	4731	4738	4746	4754	4762	4769
.52	.4777	4785	4792	4800	4808	4815	4823	4831	4839	4846
.53	.4854	4861	4869	4877	4884	4892	4900	4907	4915	4922
.54	.4930	4937	4945	4953	4960	4968	4975	4983	4990	4998
.55	.5005	5013	5020	5028	5035	5043	5050	5057	5065	5072
.56	.5080	5087	5095	5102	5109	5117	5124	5132	5139	5146
.57	.5154	5161	5168	5176	5183	5190	5198	5205	5212	5219
.58	.5227	5234	5241	5248	5256	5263	5270	5277	5285	5292
.59	.5299	5306	5313	5320	5328	5335	5342	5349	5356	5363
.60	.5370	5378	5385	5392	5399	5406	5413	5420	5427	5434
.61	.5441	5448	5455	5462	5469	5476	5483	5490	5497	5504
.62	.5511	5518	5525	5532	5539	5546	5553	5560	5567	5574
.63	.5581	5587	5594	5601	5608	5615	5622	5629	5635	5643
.64	.5649	5656	5663	5669	5676	5683	5690	5696	5703	5710
.65	.5717	5723	5730	5737	5744	5750	5757	5764	5770	5777
.66	.5784	5790	5797	5804	5810	5817	5823	5830	5837	5843
.67	.5850	5856	5863	5869	5876	5883	5889	5896	5902	5909
.68	.5915	5922	5928	5935	5941	5948	5954	5961	5967	5973
.69	.5980	5986	5993	5999	6005	6012	6018	6025	6031	6037
.70	.6044	6050	6056	6063	6069	6075	6082	6088	6094	6100
.71	.6107	6113	6119	6126	6132	6138	6144	6150	6157	6163
.72	.6169	6175	6181	6188	6194	6200	6206	6212	6218	6225
.73	.6231	6237	6243	6249	6255	6261	6267	6273	6279	6285
.74	.6291	6297	6304	6310	6316	6322	6328	6334	6340	6346
.75	.6351	6357	6363	6369	6375	6381	6387	6393	6399	6405
.76	.6411	6417	6423	6428	6434	6440	6446	6452	6458	6463
.77	.6469	6475	6481	6487	6492	6498	6504	6510	6516	6521
.78	.6527	6533	6539	6544	6550	6556	6561	6567	6573	6578
.79	.6584	6590	6595	6601	6607	6612	6618	6624	6629	6635
.80	.6640	6646	6652	6657	6663	6668	6674	6679	6685	6690
.81	.6696	6701	6707	6712	6718	6723	6729	6734	6740	6745
.82	.6751	6756	6762	6767	6772	6778	6783	6789	6794	6799
.83	.6805	6810	6815	6821	6826	6832	6837	6842	6847	6853
.84	.6858	6863	6869	6874	6879	6884	6890	6895	6900	6905
.85	.6911	6916	6921	6926	6932	6937	6942	6947	6952	6957
.86	.6963	6968	6973	6978	6983	6988	6993	6998	7004	7009
.87	.7014	7019	7024	7029	7034	7039	7044	7049	7054	7059
.88	.7064	7069	7074	7079	7084	7089	7094	7099	7104	7109
.89	.7114	7119	7124	7129	7134	7139	7143	7148	7153	7158
.90	.7163	7168	7173	7178	7182	7187	7192	7197	7202	7207
.91	.7211	7216	7221	7226	7230	7235	7240	7245	7249	7254
.92	.7259	7264	7268	7273	7278	7283	7287	7292	7297	7301
.93	.7306	7311	7315	7320	7325	7329	7334	7338	7343	7348
.94	.7352	7357	7361	7366	7371	7375	7380	7384	7389	7393
.95	.7398	7402	7407	7411	7416	7420	7425	7429	7434	7438
.96	.7443	7447	7452	7456	7461	7465	7469	7474	7478	7483
.97	.7487	7491	7496	7500	7505	7509	7513	7518	7522	7526
.98	.7531	7535	7539	7544	7548	7552	7557	7561	7565	7569
.99	.7574	7578	7582	7586	7591	7595	7599	7603	7608	7612

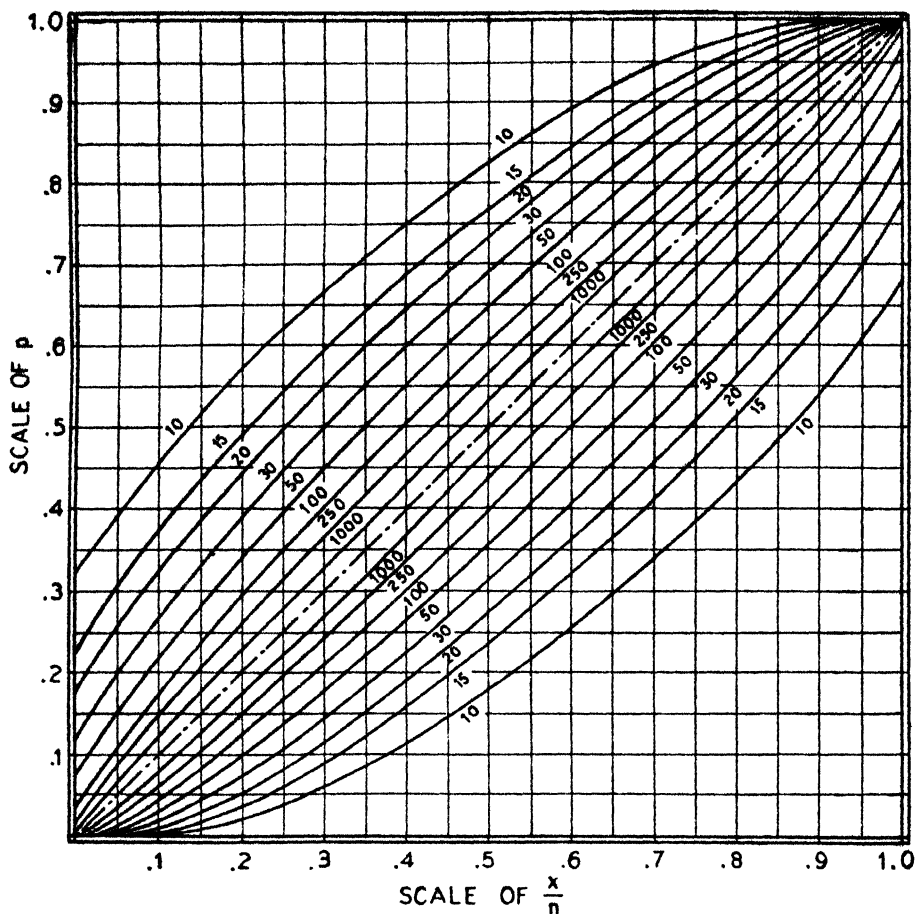
TABLE VI. VALUES OF $\tanh x'$ (cont.)

x'	0	1	2	3	4	5	6	7	8	9
1.00	.7616	7620	7624	7629	7633	7637	7641	7645	7649	7653
1.01	.7658	7662	7666	7670	7674	7678	7682	7686	7691	7695
1.02	.7699	7703	7707	7711	7715	7719	7723	7727	7731	7735
1.03	.7739	7743	7747	7751	7755	7759	7763	7767	7771	7775
1.04	.7779	7783	7787	7791	7795	7799	7802	7806	7810	7814
1.05	.7818	7822	7826	7830	7834	7837	7841	7845	7849	7853
1.06	.7857	7860	7864	7868	7872	7876	7879	7883	7887	7891
1.07	.7895	7898	7902	7906	7910	7913	7917	7921	7925	7928
1.08	.7932	7936	7939	7943	7947	7950	7954	7958	7961	7965
1.09	.7969	7972	7976	7980	7983	7987	7991	7994	7998	8001
1.10	.8005	8009	8012	8016	8019	8023	8026	8030	8034	8037
1.11	.8041	8044	8048	8051	8055	8058	8062	8065	8069	8072
1.12	.8076	8079	8083	8086	8090	8093	8096	8100	8103	8107
1.13	.8110	8114	8117	8120	8124	8127	8131	8134	8137	8141
1.14	.8144	8148	8151	8154	8158	8161	8164	8168	8171	8174
1.15	.8178	8181	8184	8187	8191	8194	8197	8201	8204	8207
1.16	.8210	8214	8217	8220	8223	8227	8230	8233	8236	8240
1.17	.8243	8246	8249	8252	8256	8259	8262	8265	8268	8271
1.18	.8275	8278	8281	8284	8287	8290	8293	8296	8300	8303
1.19	.8306	8309	8312	8315	8318	8321	8324	8327	8330	8333
1.20	.8337	8340	8343	8346	8349	8352	8355	8358	8361	8364
1.21	.8367	8370	8373	8376	8379	8382	8385	8388	8391	8394
1.22	.8397	8399	8402	8405	8408	8411	8414	8417	8420	8423
1.23	.8426	8429	8432	8434	8437	8440	8443	8446	8449	8452
1.24	.8455	8457	8460	8463	8466	8469	8472	8474	8477	8480
1.25	.8483	8486	8488	8491	8494	8497	8500	8502	8505	8508
1.26	.8511	8513	8516	8519	8522	8524	8527	8530	8533	8535
1.27	.8538	8541	8543	8546	8549	8551	8554	8557	8560	8562
1.28	.8565	8568	8570	8573	8575	8578	8581	8583	8586	8589
1.29	.8591	8594	8596	8599	8602	8604	8607	8609	8612	8615
1.30	.8617	8620	8622	8625	8627	8630	8633	8635	8638	8640
1.31	.8643	8645	8648	8650	8653	8655	8658	8660	8663	8665
1.32	.8668	8670	8673	8675	8678	8680	8683	8685	8688	8690
1.33	.8692	8695	8697	8700	8702	8705	8707	8709	8712	8714
1.34	.8717	8719	8722	8724	8726	8729	8731	8733	8736	8738
1.35	.8741	8743	8745	8748	8750	8752	8755	8757	8759	8762
1.36	.8764	8766	8769	8771	8773	8775	8778	8780	8782	8785
1.37	.8787	8789	8791	8794	8796	8798	8801	8803	8805	8807
1.38	.8810	8812	8814	8816	8818	8821	8823	8825	8827	8830
1.39	.8832	8834	8836	8838	8840	8843	8845	8847	8849	8851
1.40	.8854	8856	8858	8860	8862	8864	8866	8869	8871	8873
1.41	.8875	8877	8879	8881	8883	8886	8888	8890	8892	8894
1.42	.8896	8898	8900	8902	8904	8906	8908	8911	8913	8915
1.43	.8917	8919	8921	8923	8925	8927	8929	8931	8933	8935
1.44	.8937	8939	8941	8943	8945	8947	8949	8951	8953	8955
1.45	.8957	8959	8961	8963	8965	8967	8969	8971	8973	8975
1.46	.8977	8978	8980	8982	8984	8986	8988	8990	8992	8994
1.47	.8996	8998	9000	9001	9003	9005	9007	9009	9011	9013
1.48	.9015	9017	9018	9020	9022	9024	9026	9028	9030	9031
1.49	.9033	9035	9037	9039	9041	9042	9044	9046	9048	9050

TABLE VI. VALUES OF $\tanh z'$ (cont.)

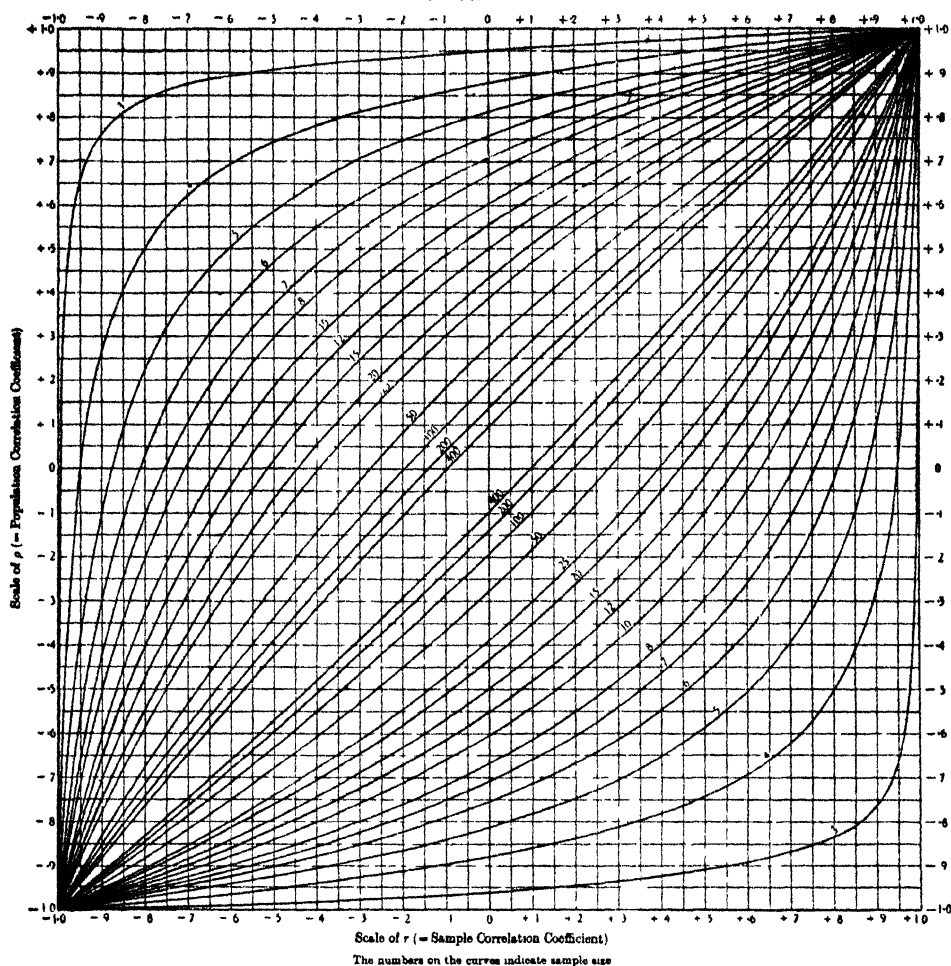
z'	0	1	2	3	4	5	6	7	8	9
1.50	.9051	9053	9055	9057	9059	9060	9062	9064	9066	9068
1.51	.9060	9071	9073	9075	9076	9078	9080	9082	9083	9085
1.52	.9067	9069	9090	9092	9094	9096	9097	9099	9101	9103
1.53	.9104	9106	9108	9109	9111	9113	9114	9116	9118	9120
1.54	.9121	9123	9125	9126	9128	9130	9131	9133	9135	9136
1.55	.9138	9140	9141	9143	9144	9146	9148	9149	9151	9153
1.56	.9154	9156	9157	9159	9161	9162	9164	9165	9167	9169
1.57	.9170	9172	9173	9175	9177	9178	9180	9181	9183	9184
1.58	.9186	9188	9189	9191	9192	9194	9195	9197	9198	9200
1.59	.9201	9203	9205	9206	9208	9209	9211	9212	9214	9215
1.60	.9217	9218	9220	9221	9223	9224	9226	9227	9229	9230
1.61	.9232	9233	9235	9236	9237	9239	9240	9242	9243	9245
1.62	.9246	9248	9249	9251	9252	9253	9255	9256	9258	9259
1.63	.9261	9262	9263	9265	9266	9268	9269	9271	9272	9273
1.64	.9275	9276	9278	9279	9280	9282	9283	9284	9286	9287
1.65	.9289	9290	9291	9293	9294	9295	9297	9298	9299	9301
1.66	.9302	9304	9305	9306	9308	9309	9310	9312	9313	9314
1.67	.9316	9317	9318	9319	9321	9322	9323	9325	9326	9327
1.68	.9329	9330	9331	9332	9334	9335	9336	9338	9339	9340
1.69	.9341	9343	9344	9345	9347	9348	9349	9350	9352	9353
1.70	.9354	9355	9357	9358	9359	9360	9362	9363	9364	9365
1.71	.9366	9368	9369	9370	9371	9373	9374	9375	9376	9377
1.72	.9379	9380	9381	9382	9383	9385	9386	9387	9388	9389
1.73	.9391	9392	9393	9394	9395	9396	9398	9399	9400	9401
1.74	.9402	9403	9405	9406	9407	9408	9409	9410	9411	9413
1.75	.9414	9415	9416	9417	9418	9419	9421	9422	9423	9424
1.76	.9425	9426	9427	9428	9429	9431	9432	9433	9434	9435
1.77	.9436	9437	9438	9439	9440	9442	9443	9444	9445	9446
1.78	.9447	9448	9449	9450	9451	9452	9453	9454	9455	9457
1.79	.9458	9459	9460	9461	9462	9463	9464	9465	9466	9467
1.8	.9468	9478	9488	9498	9508	9517	9527	9536	9545	9554
1.9	.9562	9571	9579	9587	9595	9603	9611	9618	9626	9633
2.0	.9640	9647	9654	9661	9667	9674	9680	9687	9693	9699
2.1	.9705	9710	9716	9721	9727	9732	9737	9743	9748	9753
2.2	.9757	9762	9767	9771	9776	9780	9785	9789	9793	9797
2.3	.9801	9805	9809	9812	9816	9820	9823	9827	9830	9833
2.4	.9837	9840	9843	9846	9849	9852	9855	9858	9861	9863
2.5	.9866	9869	9871	9874	9876	9879	9881	9884	9886	9888
2.6	.9890	9892	9895	9897	9899	9901	9903	9905	9906	9908
2.7	.9910	9912	9914	9915	9917	9919	9920	9922	9923	9925
2.8	.9926	9928	9929	9931	9932	9933	9935	9936	9937	9938
2.9	.9940	9941	9942	9943	9944	9945	9946	9947	9949	9950
3.	.9951	9959	9967	9973	9978	9982	9985	9988	9990	9992
4.	.9993	9995	9996	9996	9997	9998	9998	9998	9999	9999

CHART I. CONFIDENCE LIMITS (95%) FOR THE BINOMIAL DISTRIBUTION*



* This chart is reproduced with the permission of Professor E S Pearson from Clopper, C. J., and Pearson, E. S., "The use of Confidence or Fiducial Limits illustrated in the case of the Binomial," *Biometrika*, vol. 26 (1934), p. 404.

CHART II. CONFIDENCE LIMITS (95%) FOR THE CORRELATION COEFFICIENT*



* This chart is reproduced with the permission of Professor E. S. Pearson from David, F. N., *Tables of the Correlation Coefficient* The Biometrika Office, London.

Additional Answers

Page 19

1. (a) 2, (b) 4; 2. (a) 10, (b) 34.5, 44.5, etc., (c) 39.5, 39.5, etc.

Page 31

1. k .

Page 41

1. $Q_2 = 136.9$ lb., $Q_1 = 126.0$ lb., $Q_3 = 148.8$ lb., $Q = 11.4$ lb.; 2. \$5886; 6. $D = 128.6$ lb., $P = 143.8$ lb.; 7. 13.2, 99.4.

Page 61

4. 0; 9. 53; 11. 2.80 in.; 13. 68%; 14. \$5218; 15. 42.2; 16. 73; 17. 1.72%; 18. 6×10 ; 19. 4.49%; 25. $8.000 > 6.423 > 4.520$; 28. (a) 3.1 min. per problem, (b) 21.3 problems per hr.

Page 87

1. (a) 150 lbs., (b) 44; 4. 1.05%, 0.77; 6. 477.3 sec.; 145.7 sec; 10. 0.2135; 0.2659; 11. $\bar{x} = 6$, $s_x = 2$, M.A.D. = 1.5; 13. $\bar{x} = 32$, $s_x = 5$; 16. $\bar{x} = 11$, $s^2 = 10.4$; 17. $\bar{x} = 57.1$, $s = 8.75$; 18. $\bar{x} = 13$, $s = 7.21$; 22. $\bar{x}_1 = 55.7$ in., $s_1 = 1.18$ in.

Page 106

1. (a) $m_1' = 2.81$, $m_2 = 1.46$, $m_3 = -0.495$, (b) $m_1' = -0.5$, $m_2 = 1.58$, $m_3 = 0$.

Page 121

8. 25.00, 16.13, 6.72, 1.80; 10. (a) $Q_3 = 22.36$, (b) 17.64 to 22.36, (c) 24.49; 11. 63; 13. 30, 39, 45, 49, 52, 55, 58, 61, 64, 67, 70, 73, 76, 80, 86, 95; 15. $f = 257.13 \exp [(x - 69.94)^2 / 19.252]$; 17. (a) 0.06%, (b) 10.5%; 18. 793, 4207; 19. (a) 13,760, (b) 23,350, (c) 103.6%, (d) 7.42%; 20. (e) 21, 341, 780, (f) $x = 43.3, 52.5, 56.7$; 21. (b) 2314, (c) 2.28%, (d) 3.372; 22. 83, 62, 40, 17.

Page 140

9. 15; 11. 360; 21. 231; 32. $\mu = 0$, $\sigma^2 = \frac{1}{3}$, 41. $\frac{1}{2}$.

Page 158

8. (b) 0.2143, 0.1786; 9. $\mu = 5$, $\sigma = 1.826$, $\alpha_3 = 0.183$; 10. 0.00620, 0.00196; 11. (a) 0.49, (b) values of f_c are 6.8, 26.1, 37.4, 24.0, 5.8; 15. $0.487 (0.72)^2/x1$; 16. $f_c = 58.2, 29.1, 7.3, 1.4$; 17. $f_c = 108.7, 66.3, 20.2, 4.1, 0.7$.

Page 173

1. 0.919, 0.198; 3. No, $P = 0.084$; 4. No, P (one-sided) = 0.186; 5. 8.8% to 17.4%; 6. No, $(p_2 - p_1)/(\text{est. of s.d.}) = 19.1$; 7. 0.638 to 0.755; 8. No, $z = 1.2$; 9. 0.41 to 0.77; 10. No, $z = 2.42$; 11. 0.043 to 0.285; 12. $z = 1.60$, null hypothesis not rejected.

Page 194

3. 12.27 to 12.39 sec.; 6. 52 to 59; 7. $z = 2.60$, hypothesis rejected at 1% level; 8. 3.007 to 3.090. Slight bias to the right; 9. 0.84 to 2.26; 11. (a) no, (b) yes. Yields more uniform in (b); 12. No; 14. Homogeneity not rejected ($P > 0.1$); 15. $F = 1.95$ with 4 and 15 d.f.; 16. Limits for the five samples are 25.9 to 58.1, 32.8 to 71.2, 56.8 to 74.2, 36.6 to 67.0, 42.8 to 104.2, for combined sample 31.9 to 82.0. $M/c = 4.12$, no reason to doubt homogeneity; 17. No, $t = 1.09$; 18. 5.57 to 7.43, 5.08 to 7.92.

Page 216

1. $P = 0.15$ or 0.46 ; 2. $P = 0.6$, $P = 0.85$; 3. $P = 0.38$; 5. No, χ^2 (pooled) = 5.44, $P = 0.25$; 6. Yes, $P = 0.016$; 10. $\bar{R} = 12.5$, $\hat{\sigma} = 4.06$, 99% limits 0.40 to 1.68.

Page 249

2. (a) $x + 2y = 14$, $y = 3x + 3$; 3. $4x + 5y = 23$; 4. 2, 1; 7. $\sum d_i^2 = 1.12$; 11. $Y = 397 + 41.4(x - 1920)$; 15. $2v = t^2$.

Page 282

4. $\sum d_i = 0$, $\sum d_i^2 = 93.7$; 6. $s_{xy}^2 = 0.469$; 7. $\bar{x} = 125$, $\bar{y} = 80$, $s_x = 15$, $s_y = 9$, $r = 0.55$; 9. (a) $\bar{x} = 150$, $\bar{y} = 70$, $s_x = 15$, $s_y = 6$, $r = 0.25$, $Y = 0.1x + 55$, $X = 0.625y + 106.25$, (b) 71 in., 151 lb.; 11. (a) 0.044, (b) 0.36, (c) 0.74; 13. (c) 0.866, (d) 0.943; 14. No; 18. (a) yes, (b) no ($P = 0.17$), (c) no ($P = 0.33$); 19. 0.62 to 1.15; 20. 0.66 to 0.94; 21. (a) 20, 1575, 94.5, (b) 18, 1596, 88.7; 24. 2.37; 27. $Y = 0.177x + 54.8$, $\hat{\sigma}_e = 16.2$, $\hat{\sigma}_d$ indeterminate.

Page 308

2. s.s. between students = 16229 (27 d.f.), s.s. between tests = 8825 (1 d.f.), s.s. for error = 4982 (27 d.f.); 6. 0.636; 7. 0.64, agreement not significant; 11. Both non-significant; 12. $E_{yx} = E_{xy} = 0.929$, $r = 0.927$, both non-significant; 13. $\chi^2 = 76$, $C = 0.41$; 14. $Y = 1.35x + 46.8$, $r^2 = 0.482$, $E_{yx}^2 = 0.584$, non-significant; 15. yes, $\chi^2 = 6.5$; 16. $Z = 3.41x + 0.0036y + 9.1$; 17. $r_{xy \cdot z} = -0.44$, $r_{yz \cdot x} = 0.10$, $r_{xz \cdot y} = 0.76$, $r_{x \cdot yz} = 0.80$.

Page 314

12. $a_3 = 0.52$, $a_4 = 2.98$; 13. $\bar{x} = 73.2$, $s = 6.7$; 14. 70.7 to 79.3, 100, 100; 15. 214; 16. $\bar{x} = 150$, $\bar{y} = 68$, $s_x = 15$, $s_y = 2.5$, $r = 0.6$; 21. (a) 0.135%, (b) 2270; 26. $\bar{y} = 1$, $s_{ey} = 1.73$; 30. No, $z = 1.60$; 33. $\frac{2}{9}$; 34. P (one-sided) = 0.0003, highly significant; 35. For means, $t = 0.89$ (non-significant), for variances, $F = 1.48$ (non-significant); 36. s.s. between pairs = 223, s.s. between diets = 22, s.s. for error = 247.4; 37. No, $\chi^2 = 4.04$; 38. 0.479 to 0.541; 39. $a = 5.72$, $n = 1.018$; 43. yes, $P = 0.13$.

INDEX

- Aitken, A. C.**, 251
 Analysis of variance, 86, 188, 190
 parabolic regression, 294
 test scores, 288
 Arithmetic mean, *see* Mean
 Arkin, H., 31
 Array (frequency table), 270, 278
 Association
 measure of, 303
 test for, 302
 Asymptotic distribution, 212
 Attenuation, 273
 Average, 32, 42, 53
 moving, 221
- Bar** diagrams, 23, 24
 Bartlett, M. S., 190
 Bernoulli, J., 146
 Best-fitting line, 225, 226, 228
 by least squares, 228
 by moments, 226, 229
 through origin, 229
 with both variates subject to error, 279
 with equispaced data, 230
 Binomial coefficients, 144
 graph paper, 171, 174
 weights, 222
 Binomial distribution, 146, 148, 180
 approximation to, 161
 confidence limits for, 167
 fitting of, 150
 tables of, 168, 174
- Bivariate data**, 252, 269
 Bowley, A. L.
 index number, 66
 measure of skewness, 102
 Brinton, W. C., 31
 Burgess, R. W., 31
 Business time series, 242
- Calculating machines**, 2
 Calendar adjustment, 243
 Camp, B. H., 119, 123
 Carver, H. C., 123
- Charlier, L. V., 262
 check, 81, 94, 104
 Charts, 22, 23
 confidence intervals, 168, 267
 control, 192
 correlation, 273
 Chebyshev, *see* Tchebycheff
 Chi-square, 187, 197
 table of, Appendix, Table III
 test of association, 302
 test for 2×2 table, 304
 test of hypotheses, 199
 Class, 14
 boundaries, 17, 26
 interval, 14, 15
 limits, 17
 mark, 14
 Clopper, C. J., 168, 169, 174
 Cochran, W. J., 218
 Column means, 270
 Coded variate, 48
 use in calculation, 48, 62, 80, 271
 Coin-tossing, 144, 161, 163
 Column means, 273, 295
 Combinations, 130
 Compound interest, 234
 Conditional distribution, 125
 Confidence interval, 167, 168
 limits, 167, 170, 179, 183, 184, 202, 211, 265
 Contingency
 coefficient of, 303
 tables, 301, 304, 306
 Correlation, 253, 256
 index, 292
 intra-class, 316
 multiple, 307
 partial, 308
 ratios, 296
 Correlation coefficient (Pearson or product-moment), 258, 261
 confidence limits for, 267, Chart II
 for grouped variates, 268, 271
 significance of, 266
 table, 268

- Covariance, 254
 Cowden, D. J., 3, 20, 31, 240, 251
 Croxton, F. E., 3, 20, 31, 251
 Cumulative frequency polygons, 28
 Curve-fitting, 28, 90, 119
 binomial, 150
 normal, 112
 Poisson, 155
 Cycles, 220, 247
 (Slutsky-Yule effect), 223
Data, 5, 9
 classification of, 10
 David, F. N., 267, 285
 Death rates, adjusted, 71
 Deciles, 37
 Deflation (of time series), 243
 Degrees of freedom, 183
 Deming, W. E., 3, 285
 De Moivre, A., 109
 Deseasonalizing data, 243, 246
 Deviation, 47, 76
 mean absolute, 76
 standard, 60, 77, 80, 82
 standardized, 96
 sum of squares of, 84
 Dispersion, 37, 75
 relative, 83
 Distribution
 of means, 175
 of proportions, 164, 169
 of sums, 180
 Distribution function, 134
 Dixon, W. J., 4, 174, 218
 Domain (of variate), 5
 Dot diagram, *see* Scatter diagram
 Dwyer, P. S., 285
Efficiency (of estimate), 211, 212, 213
 Eisenhart, C., 202, 218, 281, 285
 Ellipse, 277
 Errors, 6
 systematic, 7
 effect of, on correlation, 273
 Estimates, consistent, 280
 from regression, 256, 262
 most-efficient, 199, 211
 unbiased, 100, 167, 182
 Events, 124, 126
 complementary, 126
 compound, 133, 136
 independent, 128
 mutually exclusive, 127
 simple, 132
 Expectation, 137, 138
 Expected value, 138
 Exponential function, 231, 234
 table, 156
 trend, 231
Factor-reversal test, 68
 Fatigue effect, 286
F-distribution, 188, 191, Appendix
 Table IV
 for ranges, 216
 Feller, W., 143
 Ferger, W. F., 63, 39
 Fisher, I., 66, 68, 74
 Fisher, Sir R. A., 78, 90, 99, 106, 205,
 218, 316
 transformation, 266
 exact method (2×2 tables), 306
 Fisher and Yates (statistical tables)
 307
 Forsyth, C. H., 251
 Fractiles, 37
 Frequencies
 absolute, 14, 17
 cumulative, 17
 cumulative relative, 111, 114
 relative, 17, 18, 91, 124
 theoretical, 114
 Frequency curves, 26, 27, 107
 distribution, 12, 14
 function, 134
 polygon, 24, 25
 Fry, T. C., 160
 Function
 integrable, 108
 linear, 224
 single-valued, 22
Galton, Sir F., 252, 255
 Gauss, C. F., 109
 inequality, 210
 Geometric mean, *see* Mean
 Gompertz curve, 241
 Goodness of fit, 197, 199
 Gosset, W. S., *see* "Student"
 Goulden, C. H., 4
 Grades (percentile ranks), 39
 Graduation, *see* Curve-fitting
 Grant, E. L., 4
 Grouping error, 47, 81, 273, 298

Growth, law of, 56, 231, 242
 g -Statistics, 101, 103

Harmonic mean, *see* Mean

Hartley, H. O., 214, 218

Haskell, A. C., 31

Histogram, 25

Hoel, P. J., 4

Hogben, L., 143

Homogeneity (of variance), 189

Homoscedastic, 266, 279

Hypergeometric distribution, 172

Hypotheses

alternative, 166

composite, 166

null, 166, 198

simple, 166

tests of, 166, 169

Index numbers, chap. V

chain linkage, 70

geometric mean, 69

harmonic mean, 69

series of, 70

Integral, definition of, 108

Interest, compound, 56

simple, 56

Interpolation, 34

Interval

class, 15

open, 15

Irregular, *see* Residual and Random

Jackson, D., 251

Jackson, R. W. B., 313

Johnson, P. O., 4

Kaplansky, I., 102, 106

Kendall, M. G., 4, 63, 205, 218

(rank correlation coefficient), 291,
313

k -Statistics, 99

Kurtosis, 102, 116, 155

Laplace, P. S., 109, 128

Laspeyres (index), 65, 67

Least squares, 225

weighted, 233

Levinson, H. C., 143

Linearity (of regression), 298

Link relatives, 237

Location, measures of, 32

Logarithmic graphs, 238, 240

Logistic, 242

Makeham curve, 241

Marginal frequency, 125

Marshall-Edgeworth (index), 66, 67

Massey, F. J., 4, 174

Mathematical model, 139, 140

Mean, arithmetic, 45, 52, 53, 54, 58

coded calculation, 48

confidence interval, 179, 183, 184

distribution of, 175

of binomial, 149, 152

of linear combination, 169

of means, 49

of Poisson, 154

of probability distribution, 136

of samples, 210

weighted, 45, 46

Mean, geometric, 54, 58

coded calculation, 62

weighted, 55, 69

Mean, harmonic, 57, 58, 69

Mean, quadratic, *see* Root mean
square

Mean absolute deviation, 76

Median, 32, 52, 53, 54

as estimate of population mean, 21

confidence limits, 211

deviation from, 77

distribution of, 211

Méré, Chevalier de, 134

Mills, F. C., 4

Mises, R. Von, 143, 201

Mitchell (index), 66, 67

Modal class, 50

Mode, 50, 52, 54, 102, 103

binomial, 148

of Poisson distribution, 154

Modified exponential curve, 240

Molina, E. C., 160

Moments, 90, 92, 93, 94, 103

about mean, 92

about origin, 91

of binomial, 150

of frequency distribution, 109, 136,
148

of ordinates, 225, 226

of Poisson, 155

standard, 98, 99

Mood, A. M., 218

Moroney, M. J., 4

Mosteller, F., 171, 174
 Moving average, simple, 221, 243
 weighted, 222
 Mudgett, B. D., 74
 Multinomial distribution, 197
 Multiple correlation, 307

Nair, K. R., 211, 218
 Neyman, J., 4, 143
 Non-parametric statistics, 197
 Nonsense correlations, 262
 Normal correlation surface, 276, 279
 Normal curve, 28, 82, 102, 111, Chap. VIII
 approximation to binomial, 161
 fitting of, 200
 history of, 109, 123
 parameters of, 110
 standard form, 110
 tables, 110, Appendix I
 Normal equations, 227, 233
 Normalized scores, 120, 121

Ogive, 29, 30, 31, 39, 40
 of normal curve, 116, 117
 on probability graph paper, 118
 Order statistics, 210
 Oscillation, *see* Cycles

Paasche (index), 65, 67
 Paired count (binomial), 172
 Paired samples, 186, 207
 Parabolic regression, 291
 Parameter, 5, 99, 148, 199
 Partial correlation, 308
 Pascal, B., 134, 145
 triangle, 145
 Payne, S. L., 20
 Pearl, R., 4, 71, 74, 242
 Pearson, E. S., 168, 169, 174
 Pearson, K., 77, 99, 252, 279
 coefficient of contingency, 303
 coefficient of correlation, 258,
 Chaps. XV, XVI
 frequency curves, 103
 measure of skewness, 101
 Percentile rank, 38, 39
 Percentiles, 37, 38
 estimation of mean by, 212
 estimation of standard deviation
 by, 213

Permutations, 129, 130
 Pie diagrams, 23
 Poisson, S. D., 153
 function, 153
 Poisson distribution, 152, 155, 157,
 160
 goodness of fit, 216
 moments, 154
 tables, 160
 Polygons, cumulative frequency, 28,
 29, 35, 38
 frequency, 24, 25
 Pooling (of class frequencies), 198
 Population, 27, 78, 90
 finite, 172, 180
 bivariate, 264
 Power functions, 238, 240
 Practice effect, 286, 288
 Price index, 64, 70
 fixed weight aggregative, 66
 ideal (I. Fisher), 66
 Price relatives, 65, 69
 Probability, 6, 27, 31, 91, 107, Chap. IX
 addition law, 128
 axioms of, 127
 classical definition of, 128
 continuous, 134
 density, 134
 graph paper (normal), 117, 118
 graph paper (binomial), 171, 174
 multiplication law, 128
 problems, 132
 Probable error, 112, 187
 Product-moment coefficient, 258
 Product notation, 44, 45
 Proportions (sample), 164
 difference of, 169, 170
 expectation of, 164
 in contingency table, 301
 variance of, 164

Quadratic mean, *see* Root mean square
 Quality control, 75, 124, 192, 204
 Quantiles, 37, 39
 Quantity index numbers, 68
 relatives, 68
 Quartiles, 35, 36
 deviation, 36
 measure of skewness, 102

- Randomness**, 138, 157, 200, 205, 281
Random numbers, 205, 207, Appendix Table V, normal, 207
Range, 75, 213
 control chart limits, 193, 194, 214
 normal population, 214
 rectangular population, 214
 quotient of, 215
Rank correlation, 289, 290
Ratio charts, 236, 237
Ratios, average of, 55
Rectangular distribution, 142, 214
Regression, 253, 255, 265
 coefficient, 254
 lines, 254, 256, 260
 parabolic, 291
 prediction from, 280
 test for linearity of, 298
Reliability (of tests), 286, 287
Residuals, 220, 228, 247
Rider, P. R., 215, 218
Rietz, H. L., 251, 285
Robbins, H., 210
Romig, H. G., 174
Root mean square, 59
Rounding off, 8, 9
Row means, 274, 295
Runs, 201
 confidence limits, 202
Run test (of sample difference), 204

Sample, 10, 78, 180
 paired, 186
 random, 10, 27
Sampling, 119
 experiment, 176
Scatter diagram, 252, 257, 261, 307
Scores, standardized, 98
Seasonal index, 244
Semi-interquartile range, 36
Semi-logarithmic graph paper, 235
Sheppard's corrections, 82, 95, 273
Sign test, 207
 table, 208
Significance, levels of, 165
 of correlation coefficients, 266
 of difference of means, 181
 of difference of proportions, 169
 of means, 177
 of proportions, 164
 of regression coefficients, 265

Significant figures, 8
Skewness, 100, 101, 102, 116
 of binomial, 150
 of Poisson distribution, 155
Slutzky, E., 223
Smith, B. Babington, 205, 218
Snedecor, G. W., 4, 191, 205, 218, 259, 285
Spearman, C., 290
Spencer's formula, 223
Split (binomial graph paper), 171
Squares, sum of, 190-192
Standard deviation, 60, 77
 coded calculation, 80
 meaning, 82
Standard error, 187
 of estimate, 267
Standard million, 71, 73
Standardized scores, 98
 variates, 96, 98, 113, 258, 260, 263
Statistic, 37, 99
Statistics, experimental, 2
 mathematical, 2
Stereogram, 276
Stratum sampling, 10
"Student", 182, 196, 265, 298
Summation notation, 42, 43
Swed, F. S., 202, 218

Tabulation, 10, 11, 16
Tchebycheff, P. L., 209
t-distribution, 182, 265
Tests, 166
 one-tailed and two-tailed, 166, 179, 186, 306
 reliability of, 286, 287
 split-half, 286
Thompson, D'Arcy W., 39
Three-variate problems, 307
Ties (in ranking), 290
Time-reversal test, 68
Time series, 219, Chap. XIV
Tippett, L. H. C., 4, 205, 218
Trend, 220, 224
 elimination of, 247
 linear, 226
Triangular distribution, 142
Tukey, J. W., 171, 174
Two-by-two table, 304, 306
 exact method, 306
 Yates' correction, 305

- Validity** (of tests), 286
Variable, 5, 21
 random, 5
Variance, 77, 84, 93
 analysis of, 86, 188, 190, 288, 294
 distribution of, 186
 estimate of, 78, 182, 183, 185
 explained and unexplained, 263
 homogeneity of, 189
 of binomial, 149, 152
 of column, 276, 296
 of estimate (regression), 262, 276
 of linear combination, 169
 of means of sub-sets, 86
 of Poisson, 154
 of probability distribution, 136
 of subsets, 89
 ratio of, 188
Variate, 5
 coded, 48, 80
 continuous, 6, 50
 discrete, 5, 35, 50
 grouped, 32, 46, 80
 independent, 261
 uncorrelated, 261
Variation, coefficient of, 84
Wald, A., 279, 280, 285
Walker, H. M., 4, 20, 123, 183, 196, 273
Walsh (index), 66
Weatherburn, C. E., 4
Weighted least squares, 233, 273
 mean, 45, 55, 69
Wilks, S. S., 4, 78, 218
Wilson, E. B., 300, 313
Winsor, C. P., 281, 285
Wolfenden, H. H., 4
Yates, F., 205, 218
 correction in 2×2 tables, 305
Yule, G. U., 4, 63, 223, 262





DELHI POLYTECHNIC
LIBRARY

CLASS NO. 311

BOOK NO. W 47 B

ACCESSION NO. 8616